

**Data mining using open source software for small business  
Including case study.**

Antoine Dubuis

<b>Authors</b> Antoine Dubuis	
<b>The title of your thesis</b> Data mining using open source software for small business	<b>Number of pages and appendices</b> 63+4
<b>Supervisors</b> Arvo Lipitsainen – Advisor from Haaga-Helia University of Applied Sciences Prof. Dr. Michael Schumacher – Advisor from HES-SO University of Applied Sciences	
<p>Data mining is a known as the process of discovering patterns on large datasets. It is nowadays really popular technologies used by multinational firms to have a deep understanding of their business and customers. An entreprise like Amazon have gained considerable assets with it.</p> <p>This thesis aims to answer the question: “Is the data mining also available for small companies?”</p> <p>In order to have a clear solution, the initial step is to present this environment in a comprehensive way. During the first part, all knowledge needed to have a good insight of that field will be explained.</p> <p>Then the second part is a case study on a wine-selling business called “La cave du Palais de Justice”. This research will be based on their accounting data and the software used for this task will be R. This case will pass through all phases of the process.</p> <p>At the end, we will discuss the result of the study as well as the problem we faced and their solutions. Then we will argue whether data mining and their open-source system are available for small business.</p>	
<b>Key words</b> Data mining, data, Knowledge discovery in databases, KDD, small business, Statistics, R, Time series, Association rules, MBA, Market Basket Analysis	

## Table of contents

Illustrations.....	iv
Table of abbreviations .....	v
<b>1 Background.....</b>	<b>2</b>
1.1 Thesis significance.....	2
1.2 Research Problem .....	2
1.3 Thesis Structure.....	3
1.4 Limits and risks.....	3
<b>2 Data mining .....</b>	<b>4</b>
2.1 Introduction .....	4
2.2 Data, Information, Knowledge, Wisdom (DIKW) Hierarchy .....	5
2.3 What kind of data are we collecting?.....	6
2.4 Knowledge discovery in databases process .....	8
2.4.1 Data selection.....	8
2.4.2 Data Pre-Processing.....	9
2.4.3 Data Transformation .....	9
2.4.4 Data mining.....	10
2.4.5 Pattern Evaluation.....	10
2.4.6 Knowledge Presentation .....	11
2.5 What can be discovered? .....	11
2.5.1 Classification .....	11
2.5.2 Regression .....	12
2.5.3 Association Rules .....	12
2.5.4 Clustering analysis .....	13
2.5.5 Times Series .....	14
<b>3 Case study .....</b>	<b>15</b>
3.1 Case description.....	15
3.2 Case Objectives .....	15
3.3 Software choices.....	16
3.4 Data available.....	16
3.5 Case Study process.....	17
3.5.1 Data Selection .....	17

3.5.2	Data Pre-Processing.....	18
3.5.2.1	Data Import .....	18
3.5.2.2	Selecting variable .....	19
3.5.2.3	Data cleaning.....	20
3.5.3	Data transformation.....	27
3.5.3.1	Discretize Variables.....	27
3.5.3.2	Construct new variables .....	28
3.5.4	Data mining.....	28
3.5.5	Mining Customer Behaviour .....	28
3.5.5.1	Time Series Forecasting.....	39
3.5.6	Knowledge Presentation .....	43
<b>4</b>	<b>Discussion.....</b>	<b>46</b>
4.1	Result of the case study .....	46
4.2	Problems Encountered.....	47
4.2.1	Data Quality .....	47
4.2.2	Lack of Dimension.....	49
4.2.3	Communication .....	50
4.2.4	Time .....	51
4.3	Is datamining available for small companies? .....	51
4.4	Is open source data mining a viable choice? .....	54
4.5	Evaluation of own-learning .....	56
4.6	Methodology .....	57
<b>5</b>	<b>Conclusion .....</b>	<b>58</b>
<b>6</b>	<b>Bibliography .....</b>	<b>59</b>
	<b>Appendices .....</b>	<b>A</b>
	Appendix A: Tableau dashboard: 2016 Sales forecast per month.....	A
	Appendix B: Tableau dashboard: Sales per type, country and region.....	B
	Appendix C: Tableau dashboard: Margin analysis .....	C
	Appendix D: Tableau dashboard: Average sales per day of week and hours .....	D

## Acknowledgements

First, I would like to thank my thesis advisor in Finland, Mr. Arvo Lipitsainen for his suggestions and help during this project as well as my supervisor in Switzerland Mr. Michael Schumacher.

My sincere appreciation goes naturally to Mr. Juhani Välimäki, The Move Office and Mr. David Wannier, Head of BIT department of HES-SO, Valais/Wallis, who made this international exchange possible.

I am also thanking Mr. Valentin Berclaz for his support, his valuable inputs and peer reviewing this thesis.

Finally, I would like to thank Dynaxis Sàrl who gave me the opportunity to make an internship in that field. Thanks also to "La Cave du Palais de Justice" who share me their data.

## Illustrations

Figure 1: DIWK Pyramid (High Need 2 Know, 2016) .....	5
Figure 2: KDD process (Rithme, 2016).....	8
Figure 3: Summary on articles .....	21
Figure 4: Summary of semantic errors .....	23
Figure 5 Boxplot of margin percentage with and without extreme values .....	23
Figure 6: Summary of group and country columns .....	24
Figure 7: Summary of the old pattern rows.....	24
Figure 8: Table showing the alc_type values .....	26
Figure 9: Table showing the alc_type final values .....	26
Figure 10: Result of the price discretization.....	27
Figure 11: Summary of the transaction object.....	31
Figure 12: Item repartition on the transaction set for the first 10000 and all the set .....	32
Figure 13: Summary on objects rules .....	33
Figure 14: Inspect rules by support .....	34
Figure 15: Inspect rules by confidence .....	34
Figure 16: Top 30 frequency of transaction set variables.....	36
Figure 17: Summary of the rules .....	37
Figure 18: Inspect pruned rules.....	37
Figure 19: Inspect rules with alcohol type as consequent .....	37
Figure 20: Inspect rules with alcohol type as consequent without VRG.....	38
Figure 21: Inspect rules with June or July as antecedent.....	38
Figure 22: ArulesViz graph of the 10 first rules.....	38
Figure 23: Time series output.....	40
Figure 24: Product sales time series.....	40
Figure 25: Times series components .....	40
Figure 26: auto.arima trace.....	42
Figure 27: Forecast output.....	42
Figure 28 Sales forecast using ARIMA .....	43
Figure 29: Dashboard of products, origin and type .....	45
Figure 30: ARIMA forecast on Tableau .....	45
Figure 31: Gartner 2016 MQ for Advanced Analytics Platforms (KDnuggets, 2016) .....	54

## Table of abbreviations

<b>AOC</b>	(Swiss label) controlled designation of origin
<b>ARIMA</b>	AutoRegressive Integrated Moving Average
<b>ARULES</b>	Association rules
<b>BI</b>	Business Intelligence
<b>CLP</b>	Customer Loyalty Program
<b>CPJ</b>	Cave Palais de Justice
<b>CRISP-DM</b>	Cross-Industry Standard Process for Data Mining
<b>CSV</b>	Comma Separated Value
<b>DBA</b>	Database administrator
<b>DIKW</b>	Data Information Knowledge Wisdom
<b>DM</b>	Data mining
<b>GPS</b>	Global Positioning System
<b>IBM</b>	International Business Machines Corporation
<b>IS</b>	Information System
<b>ISO</b>	International Organization for Standardization
<b>IT</b>	Information technology
<b>KDD</b>	Knowledge Discovery in Databases
<b>LHS</b>	Left-hand side
<b>MQ</b>	Magic Quadrant
<b>RHS</b>	Right-hand side
<b>SAP</b>	Systems, Applications and Products
<b>Sàrl</b>	(French abbreviation) Limited responsibility company
<b>TS</b>	Times series
<b>VBL</b>	(French) White wine
<b>VRG</b>	(French) Red wine
<b>WWW</b>	World Wide Web
<b>XL</b>	Excel

# 1 Background

Dynaxis Sàrl, a small business intelligence company from Switzerland, want to provide her customer data mining service in order to enlarge their offer. This BI Company is working along with a lot of software partners such as SAP, Microstrategy, Oracle or IBM.

Actually, in Switzerland more than two thirds of the companies are small (less than 249 employees) and most of them cannot afford services such as SAP.

That is why they want to explore the solutions offered by open source applications and freeware.

The advantage of this thesis for this company would be

- To understand how data mining is working
- Discover open source solutions
- Realize a project for one customer

## 1.1 Thesis significance

The findings of this thesis will be helpful for the small companies who want to start using data mining process to have a better understanding of their business. They should be able to learn the basics of data mining, have an overview of the affordable software and have some tips to increase the success of their studies.

This document is also going to be a good guidance for the IT community who want to have a simplified insight of this field. This research is also relevant for Dynaxis Sàrl because it will help them to offer new services to their customer as well as having a better understanding of the opportunities offered by data mining technologies. Lastly, an extern company will also benefit from the case study. It will permit them to have a data mining insight of their company.

## 1.2 Research Problem

There are multiple objectives for this thesis. This document should be able to answer the following questions:

- Is open source software a viable solution for data mining?
- Are small businesses mature enough for data mining?



### **1.3 Thesis Structure**

The first part of this document is a theoretical explanation of data mining. It will introduce in an understandable way the goal, the available data and processes of that technology. It will also contain a description of the different pattern that data mining can discover.

The second section will be focused a concrete case study. It will cover all the Knowledge Discovery in Database process. This study will point out difficulties related to data mining for a small business. After that demonstration, we will explain the problem that has arisen from it and present some solutions to avoid them.

We will conclude this thesis by looking how usable is open source software and if the small company are ready for it.

### **1.4 Limits and risks**

The theoretical part will not contain the entire mathematical component related to data mining because this section is supposed to be understandable for IT workers. It might be difficult to write a complete explanation on this topic without speaking about the statistical point of view. This can be a threat to over simplify these explanations.

The lack of data and dimension are a huge risk for this applied research. It can cause the project to fail because we cannot produce interesting predictive models without having enough data to observe.

The case study will be made for a small company. Therefore, we can guess that they do not possess a centralized database, which means that the entries won't be pre-processed. This is a problem because data quality is a big issue for that kind of project. The time available for that research can also have a big impact if there is some unexpected problem arising.

## 2 Data mining

### 2.1 Introduction

The amount of data created and available has been growing exponentially during the last decades. News websites, such as Science Daily, were reporting that 90 % of the world data has been generated over the last two years (Science Daily, 2013). This colossal number of measurements come from everywhere, from sensors, posts on social media to GPS signals. In the business field, companies are also collecting various kinds of data such as customer information, sales or financial data.

All of this can be really valuable for an enterprise because it can help them to have a competitive advantage on the market. But this amount of data is now more and more difficult to analyse due to his size.

As the author of Megatrends John Naisbitt said, ‘We are drowning in information but starved for knowledge’ (Naisbitt, 1982). This sentence is perfectly explaining our decade problem. Companies need to find a way to keep an eye on their business. A solution to that problem was invented in the nineties, data mining.

It could be designate as the method of discovering patterns in large datasets. This is a part of the Knowledge Discovery in databases (KDD). Which is the process of analysing raw data from different sources and to turn it into valuable information (Technopedia, 2016). The definition of this technology is not precise because it had to encapsulate multiple fields covered by data mining, which are statistics, artificial intelligence, databases, data analysis and much more. (Britannica, 2016)

Initially, the exclusive target of data mining was well-structured data, such as relational databases (Nestorov, 2000). An example data mining using structured data is the mighty story of beer and diaper on Wal-Mart:

Some time ago, Wal-Mart decided to combine the data from its loyalty card system with that from its point of sale systems. The former provided Wal-Mart with demographic data about its customers, the latter told it where, when and what those customers bought. Once combined, the data was mined extensively and many correlations appeared.

On Friday afternoons, young American males who buy diapers also have the predisposition to buy beer. No one had predicted that result, so no one would ever have even asked the question in the first place. Hence, this is an excellent example of the difference between data mining and querying. (The Register, 2006)

Nowadays data mining is not anymore used only with well-designed data. With World Wide Web (WWW) and social media explosion, there are plenty of new data sources available that are hiding more information than ever. And this field can no longer be secluded to structured data.

The types of data have changed a lot these past years, images, video, and free text are now available for us to analyse. Companies such as IBM are already sure that data mining will be one of the most important technologies in the next decades.

## 2.2 Data, Information, Knowledge, Wisdom (DIKW) Hierarchy

Before starting to speak about data mining, it is important to explain the difference between Data, Information, Knowledge and Wisdom. The Literature is representing this hierarchy with the help of a four-levelled pyramid. The Figure 1 below is illustrating this pyramid.

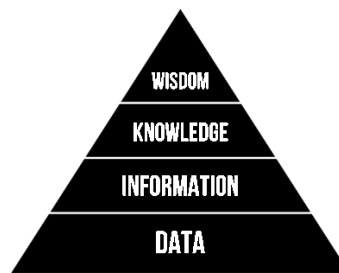


Figure 1: DIWK Pyramid (High Need 2 Know, 2016)

This pyramid describes a hierarchical relationship between these terms. Under we will explain all the DIKW layers using an everyday concrete example.

**Data** can be defined a fact such as a text or a number that can be processed by a machine. Without interpretation or computation data has no meaning.

*Red Is data*

**Information** is organized data that have a context and a meaning. This layer represents processed data and it is the base of the knowledge. The information provides answers to “who what where when” questions.

*Traffic light is red is information*

**Knowledge** is the synthesis of multiple information sources. It should be able to answer the question, ‘Why.’

*The traffic light in front of me is red*

**Wisdom** is a concept related to the ability to use knowledge to make a good judgement. It is often dismissed from the hierarchy because it is a “non-material” layer (Complexity Media, 2015).

*I should stop the car*

### 2.3 What kind of data are we collecting?

Today it's hard to think of an everyday task that is not creating data. Indeed, data collection and creation are involved in our everyday-life. From the simple numerical measurement to your credit card usage, data are surrounding us. In this section we're going to look at different kinds of data that we are currently collecting.

The Computing Science department of Alberta provided us this list of information collected in the database.

- **Business Transactions:**

Every transaction in the business field is most of the time store for perpetuity. All of them are time-related and are often linked to an extern entity such as clients, products or companies. The example of the supermarket shows us that every day there is terabytes of data collected describing the basket, article, and customers. Storing this data is not a problem thanks to the hard drive's prices but analysis of this data is definitely an important concern for businesses who are struggling to survive in such competitive area (Zaïane, 1999).

- **Scientific Data**

Research fields are also gathering a myriad of measurement such as weather stations collecting every second all variables related to humidity, CO<sub>2</sub>, and temperatures. All scientific departments are producing a lot of information with these measurements and sensors. Organization such as Scientific Data are giving open-access to multiple datasets in order to maximize the reuse and discovery (Scientific Data, 2016).

- **Medical and Personal Data**

From government to customer files, there is a very large collection of data that are concerning us. All these entities are gathering a gigantic amount of personal data that can be used to have a better understanding of the customer behaviour.

Research found on Technology science tell us the following : “73% of Android apps shared personal information such as email address with third parties, and 47% of iOS apps shared geo-coordinate and other location data with third parties”. (Jinyan Zang and Cie, 2015) As we see these data exist and are collected by our everyday application.

But there is, of course, data protection directive that checks if the process and use are legal or not. More Information can be found on the European commission website.

- **Surveillance Video and picture**

Thanks to the collapse of both storage and video camera prices, we can now digitalize these records and store them for statistical purpose (Zaïane, 1999).

- **Games and sport**

Our society is passionate by sport and competition. If we take a challenging game like soccer, we can easily see that every player, club, tournament, stadium is described by a colossal amount of information. All this information is interesting for fans and journalists but it can also be used by the manager to try to guess the next scores. There are already websites that are using data mining process to forecast football result. An example of it is the company Betegy who claim to predict football score, using self-learning algorithm and data mining, with 75 % accuracy (Betegy, 2016).

- **Digital Media (Video, Photos)**

With the democratization of the camera and smartphone, there is a huge boom on website such as YouTube. Every minute, there are 72 new hours of video on the Google platform, more than 350 000 pictures shared on WhatsApp and 216 000 on Instagram. All of these are, of course, stored in a server. (Business Insider, 2015)

- **Social Media**

At the end of 2015, there was according to the website Statista, 1.3 billion active users on Facebook (Statista, 2015). All the Facebook users are posting what they like by sharing content, text or images. All this information is really valuable to understand how people are reacting to something. Twitter and other social media are also providing colossal amount of data that can be used to analyse trends of objects.

- **World Wide Web**

The World Wide Web is the biggest repository of data ever built. It is composed of billions documents and many different formats. Its size is constantly increasing and the content is really complex due to a large amount of format.

## 2.4 Knowledge discovery in databases process

Data mining, also called Knowledge discovery in databases, is the application of statistics, artificial intelligence machine learning, and database exploration in order to extract unknown useful information.

People are often using the word data mining as a synonym to KDD but as we see in the figure below, data mining is, in fact, an iterative step of the Knowledge discovery process. (Han & Kamber, 2011) This process, like you see in the Figure 2 below, is composed of many steps that we are going to explain in this section

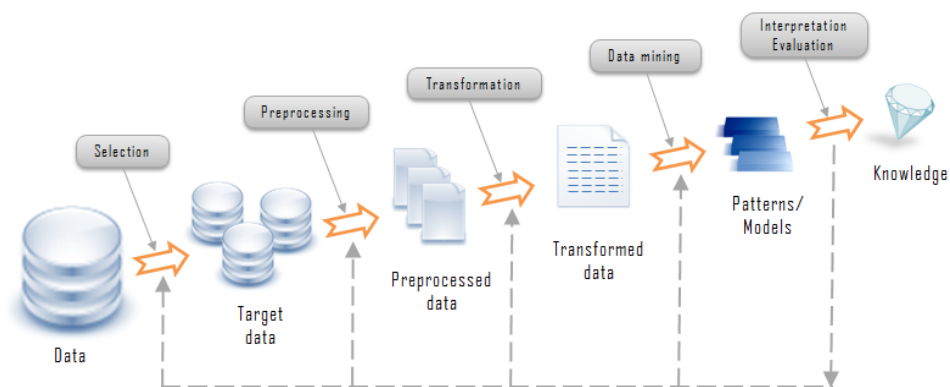


Figure 2: KDD process (Rithme, 2016)

### 2.4.1 Data selection

The first phase of a KDD process involves the selection of data variables. Data selection is really depending on the goal of the analysis as well as the variable importance in the business field. Having a good understanding of the business environment is an important asset for this phase. It's important to note that this sub-process deal with both, variables and record selections (Tuffery, 2015).

### 2.4.2 Data Pre-Processing

Data Pre-Processing, also called data cleaning, is the second step of the process. It deals with detecting and removing error from the data to increase the quality of the dataset. Data quality problems such as hand-written data, missing information or semantically invalid data are often present in a dataset. When data come from different sources, data cleaning needs will increase significantly because there is more probability that there are redundant data and different formatting. (Rahm, 2016)

This step is really important because if the dataset contains missing information then the next steps will produce bad statistics. The next steps cannot work correctly without a really good cleaning. This phase description can be summed up with this computer science term 'Garbage in, Garbage out' (GIGO).

### 2.4.3 Data Transformation

This section is presenting the data transformation phase which is used to convert data from a source format to a destination format (Badie, 2016). The goal of this phase is to transform the outcome of the pre-processing step to a dataset appropriate for data mining.

Multiple kinds of transformation can be used to reach that objectives such as attribute construction, Normalization, smoothing or Aggregation. (Badie, 2016) There is below a basic explanation for each of these techniques.

**Attribute construction** is used to construct and add values a dataset. For example, birth data can be derivate to obtain the variable age.

**Normalization** or standardization is used to adjust the data for a variable by reducing it ranges for example  $[0, 1]$  (Trendowicz, 2012).

**Smoothing** is used to remove statistical noise by capturing only the important pattern (vcefurthemaths, 2011). To do this, we often use a moving average, logarithms, and other mathematical function. The goal is to make the data follow a linear line.

**Aggregation** is used to create sum or average of the data point available. It is used when we want to change the granularity of the data set. (Trendowicz, 2012) For example, sum the sales per day instead of every sales entry.

#### 2.4.4 Data mining

Now that we have described the all the previous step, it's time to explain the heart of this process, his objective and result. Like the explanation on the introduction, data mining is the automatic extraction of unknown and useful information from data. (Frank, 2016) The goal of data mining is to find useful patterns. This can be subdivided into two different sub-fields.

The first one is **descriptive** data mining or **unsupervised**. His goal is to find patterns within the object by looking at their variable. Association rules, for example, aim to find patterns between the different variable factors. Other patterns such as clustering are regrouping similar row according to variable similarities (Oracle, 2016). The second one is called **predictive** data mining or **supervised**. In this field, we are looking to predict a target variable based on other variables values. An example is the pattern of customers that have churned. Once we statistically know, based on the pattern, which type of customers is more likely to churn, we can predict if the client x is going to switch his loyalty as well. Then we can make a marketing action to avoid it. Analysis of forecast is also a member of the predictive analysis because in this case we aim to find future values based on the previous one.

#### 2.4.5 Pattern Evaluation

A data mining analysis may generate dozens of patterns, but some of them are not interesting. The pattern evaluation is there to evaluate if the findings are valuable or not. There are multiple points that need to be checked to validate a pattern.

A pattern that is describing a relation that is obvious and already known is useless. The first criteria are the accuracy of the predictive model. The correctness of the prediction depends on the business field. In the marketing, a score higher than the randomness is usable. Contrariwise, in the medical area, a minimum of 95% is needed. (Tuffery, 2015) Secondly, the reliability of the model needs to be evaluated. The models need to be checked with some testing data to ensure that it can be reused afterwards.



### 2.4.6 Knowledge Presentation

This is the final phase of the process. The discovered knowledge is visually represented to the user. The goal is to show the correlation or significance of data. Data visualizations go further than the simple Excel graphs. Professional tools such as Tableau, MicroStrategy or JavaScript libraries are used in this step. Languages like R are able to create plot natively or better-looking plot with the help of packages such as ggplot2 or lattice.

## 2.5 What can be discovered?

A pattern in data mining is a trend or a structure that is on the dataset but invisible without processing. When we are registering data in the database, some regularity, similarity or outlier occurs. And those are creating patterns. (UCLA Anderson, 2016) For Example, in basket analysis, we can see that people who are buying gin are constantly buying tonic. This is a well-known pattern but there exist most probably a lot of unknown patterns like in the beer and diaper quote. The kind of pattern discovered depend on the data mining task we used which can be either descriptive or predictive.

Descriptive functions are used to describe the relationship between factors. On the other hand, predictive data mining is used to predict a variable based on the available data. The difference could be sum with this sentence: "Predictive modelling is all about "what is likely to happen?", whereas explanatory modelling is all about "what can we do about it?" " (probabilityislogic, 2011). This chapter aims to explain the most popular pattern that can be found into datasets.

### 2.5.1 Classification

Classification model is used to predict a categorical variable. The methodology is to find variables that are highly correlated to the variable we want to predict and then to create a tree to assign that variable (Tutorials point, 2016) .

A concrete example of the classification is the credit scoring. Banks are using classification to know whether to loan money or not based on the person variable such as age, employment, origin...

In order to predict the output, the model needs to be trained with a sample of the dataset containing the attribute with the predicted outcome. The algorithm is trying to discover relationship between the attribute that would make possible to predict the outcome. Then if the result is accurate enough, the model can be used on other datasets with empty target value. (Tutorials point, 2016)

The simplest algorithm for classification is the decision tree. It product a set of branching decision that ends in a classification. Another algorithm such as the nearest neighbour can be used for this task.

### 2.5.2 Regression

Regression is similar to classification except that this is used when the target variable is a numerical value. Regression is used to analyse relation between one variable and many others. An example of regression usage is to calculate a salary based on other variables. The regression process remains the same as the classification one.

### 2.5.3 Association Rules

Association rules are belonging to the descriptive subgroup. The goal of this process is to find rules and relation between the different values.

This method is most of the time-related use in Market basket analysis. His goal, in that example is to analyse the behaviour of the customer. These analyses can be used to find unknown relation between to produce on the shop. This information can later be used for marketing purpose like optimizing the product location, create enhanced promotion... This pattern will be explained by the school example below.

Basket ID	Milk	Bread	Butter	Beer
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

The above table is containing supermarket tickets. In Association rules, this table is named transaction matrix, the row are representing the transaction and the columns represent items. For example, the Basket 5 contains only bread. Every entry is then transform into bit vector.

We would like to find the probability that the customer who is buying bread and milk are also buying butter. This association can be represented as the following

$$\{\mathbf{Bread \& Milk}\} \rightarrow \{\mathbf{Butter}\}$$

Every rule has two distinct parts, **the antecedent and the consequent**.

The antecedent is the conditional part, here it is **{Bread & Milk}**.

The consequent is the implies of the antecedent, here it is **{Butter}**.

The first indicator we can calculate is the support. It is the probability the couple antecedent & consequent on the dataset (IT Miner, 2015).

$$\mathbf{S = Support = Frequency (\{Bread \& Milk \& Butter\}) / Total Transaction}$$

In that dataset, the rules **{Bread & Milk} -> {Butter}** is occurring 20% of the time.

On the two clients that are buying the antecedent **{Bread & Milk}**, 50% is also implying the consequent **{Butter}**. This second percentage is the confidence. That indicator is measuring the probability that the consequent is true, knowing the antecedent.

$$\mathbf{Confidence = FRQ(\{Bread \& Milk \& Butter\}) / FRQ(\{Bread \& Milk\})}$$

The last indicator useful on the association rules analysis is the **Lift**. It allows us to measure the predictive capacity of the rules. It looks if the occurrence of both antecedent and consequent are more likely to be higher when combined. When the lift is greater than 1, the rules have a significant predictive value (IT Miner, 2015).

$$\mathbf{Lift = S (\{Bread \& Milk \& Butter\}) / (S \{Pain \& Lait\} * S \{Beurre\})}$$

#### 2.5.4 Clustering analysis

Clustering analysis groups objects based on information found in the data describing the objects (University of Minnesota, 2000). His goal is to regroup similar objects into a group called cluster. The similarity within the cluster and the difference between them are two important assets to have a distinctive clustering analysis.

This analysis belongs to unsupervised learning or descriptive data mining because we are not searching to predict any values. Clustering algorithms is used when there are no groupings on the dataset.

### 2.5.5 Times Series

A time series is a list of numerical value that represents the evolution of quantity over time. This dataset can be interpreted to understand the past behaviour and predict the future trend. It can be represented on a graph where the y axis represents the values to analyse and the x representing the regular period that can be years, quarter, months and even seconds (Pollack, 2010). These lists of values are a combination of the following four components:

- **Trend:** it is the orientation of the signal, Trend is telling the global behaviour of the dataset, if the values are increasing or decreasing.
- **Cycles:** Cycles are phenomena that happened at least 2 times on the time series. There are two types of cycle.
  - o **Periodic cycle:** It is a defined by the number of times it is covering. For the Easter eggs, the phenomenon is periodic because the date is changing every year but the period still has the same length. (Singhal, 2010)
  - o **Seasonal cycle:** These are short-term variation depending on a date or an hour. A good example explaining this term could be the season for a ski resort.
- **Fluctuation:** When we subtract from the signal the trend and cycle components, we end up with the fluctuation. They are irregular variation tat at short in duration and have no regularity in the occurrence pattern (Singhal, 2010). These special events could come from unexpected events such as war, earthquakes or viruses.
- **Irregular factors:** Once we are taking everything of the signal, we end up with the random component that is unpredictable.

The goal of this analysis is to make a forecast of the future values based on the component we describe before. Algorithms are separately calculating the trends, cycle and seasonality. Trend calculation can be easily done with a linear regression. This will give us the function of the trend in this format  $AX+B = Y$ .

Concerning the calculation of the seasonality effect, we need to calculate as many indicators as data points on the season. For a quarter-time series, we need to calculate what impact has the four quarters on the values. (Pollack, 2010) Then we know that in the 4<sup>th</sup> quarter, the values are 10% percent higher than average and this is helpful

to predict values. Then the rest of the value can be interpreted as fluctuations or irregular factor according to their presence on the dataset.

### **3 Case study**

#### **3.1 Case description**

This study cases concern La Cave du Palais de Justice located in Geneva. This is a company organized as a limited company. His business goal is to sell wine and brandy. This shop own two stores. The first one is located in Geneva (Place du Bourg-de-four 1, 1204 Geneva, Switzerland) and the second site is on the town airport. This company provide a lot of products that come from various places such as Europa Africa and America. The available products are ranged from the casual wine to the luxurious brandy. This company is quite small, therefore these customers don't have an IT infrastructure except their accounting software. This system is WinBIZ and it is installed on a Windows XP virtual machine.

WinBIZ is company management software that can be used to centralized accounting, billing and time tracking systems for small and medium-size companies. This system is developed in Martigny, Switzerland. (WinBIZ, 2016)

The manager of this company has presented us the administrator of that system. And this person didn't want us to have a direct access to the database, so we agree to receive the data as flat files. The goal of this project for Dynaxis is to see if data mining can be offered to small companies that don't have a solid IT infrastructure and don't want to invest in expensive solutions such as SPSS or SAP Predictive analytics.

For the customer, the goal is obviously to obtain new information and analysis to have a better understanding of his business.

#### **3.2 Case Objectives**

Our customer was excited about association rules because he wanted to know better what was his client buying together, his goal was to discover some relationship between product in order to perform cross selling or optimize gift packages. Secondly, our client told us that he would like to have a product sales forecast for the next years. We have also agreed to create a dashboard so he could rapidly visualize his

company data because they don't have they do not own a Business Intelligence solutions, plus it is not possible at the moment to have a good visualization because of the data quality.

### **3.3 Software choices**

As written in the study case description, the CPJ (Cave Palais de Justice) is a small-sized company. Therefore, they want to see the concrete data mining benefits without investing a lot of money. They cannot simply afford business solutions such as SAP. The administrator would like to use open software in order to reduce software fees while having a tool that has all functionalities.

For this project, we first chose KNIME because it is really easy to use and people can rapidly learn how to use it. This open-source application allows us to build a node by nodes data mining process. The process is easy to understand because of the graphical representation of the nodes. In addition KNIME can easily integrate data from multiple sources. It does not require any programming skills to have a basic use of this software. This software is also really well graded on the Gartner 2016 Magic Quadrant for Advanced Analytics Platforms. (KDnuggets, 2016)

However the lack of documentation and community has made his usage really hard. As a matter of fact, answers to problems aren't easy to found on the Web. The alternative is the R language and statistical environment that allows us to perform data mining. This option is more difficult compared to KNIME but R have an enormous community on the Web. On the Stackoverflow website, R is 36th most popular tags with more than 120'000 questions. (stackoverflow, 2016) We have therefore chosen R for this project because of this reason.

### **3.4 Data available**

The first step of this project was to contact the Accounting system's administrator. He first presented us their IT-Infrastructure. As explained in the description, we were not allowed to have a direct access to the system database because this study was a pilot project. We agreed to receive Excel files containing the data.

For this project we have first receive the sales data from years 2010 to 2015. After that (two weeks later), we received the article dataset and the database's field description.

### 3.5 Case Study process

For this case study we will use the Knowledge discovery process that we presented earlier on this document. We will first start by preparing the data, so we could use them to discover some patterns. Each step will be explained with the help of R snippet and function output. We aim to describe that complex process with the appropriate words.

#### 3.5.1 Data Selection

The files coming from winBIZ were containing all fields concerning sales and articles. These variables represented 224 columns for the sales dataset.

At the beginning of the project, we didn't receive the data model including the description of each variable so we first chose to delete

- Columns with a high missing value rate
- Duplicated columns
- Unique value variables
- Empty columns

After this first phase, we named the columns based on the observation on the dataset. We realized, for example that some variable such as the client information were not usable because most of the customer are occasional customers that are paying by cash. We have deduced from it, that the fields about the type of selling and billing information were not usable. With the data selection sub-process we reduced the sale variables from 224 to 28 variables.

In order to automate the data selection we have built an Excel file containing the column names we selected. This file permitted us to create a mapping table to rename the variable and to comment the variables. You see below a sample of this Excel files. The first two columns are acting like mapping table.

<b>Id_column</b>	<b>Text value</b>	<b>desc</b>
ar_numero	dl_article	Article id

For this mapping table, we preferred to use Excel format over the common CSV because we want this document to be easily editable. We have also added some layout to make it more comfortable to use.

### 3.5.2 Data Pre-Processing

This phase is covering data importation and data cleaning. Here we will explain the tasks we have achieved to improve the quality of the document. This phase is representing more than one fifth of the time allocated to a data mining project.

#### 3.5.2.1 Data Import

The first part of data pre-processing is to import the dataset and merge them together in R software. The result is a unique dataset containing all the merged dataset. R allows us to manage import from multiple kinds of dataset. R is handling all databases, flat files and Web import. But in our case, data were delivered as Excel files. It's important to note that we chose for both storage and speed purposes to transform Excel files into CSV. You find under the sample of code concerning this step.

```
#get all the files names
files <- list.files(path = 'data/sales/year')

#loop through all of them
for(x in 1:length(files)){
  #store the path files
  path<-c('data','sales','year', files[x])
  #read csv at this location
  temp <- read.csv(file = paste(path, collapse = '/') ,
                   header = T, sep = ';',
                   na.strings = '?')
  #if it's the first then we create Sales_Master
  if(x == 1){ Sales_Master <- temp }

  #Otherwise we just attach the data to sales Master
  else{ Sales_Master <- rbind(Sales_Master, temp)}
}
```

In this R snippet, we are first getting all the filename located in the year repository. Then we are looping through that entire list to read these files and add them to the master file. We can notice the file organization on the project. We chose this implementation because we want to make the automation as easy as possible. The administrators only need to store the csv files on the year repository to merge it with the other dataset.

```
CAVEPALAISJUSTICE\DATA
\---sales
|   Sales.csv
|   Variables.xlsx
|
\---year
    data_2010.csv
```



### 3.5.2.2 Selecting variable

Now that we have a single dataset for both sales and articles data, we can apply them the result of the data selection phase. Under, there is a code snippet concerning the variable selection of articles.

```
## Selecting attributes
#####

#loading the excel containing variables and explanations
variables<-readWorksheet(
  loadWorkbook('data/article/Variables_article.xlsx'),
  sheet=1)

# get column to filter
col_name <- variables[,1]

#filter the columns
articles <- articles[, (names(articles) %in% col_name)]

# get column names from xls
col_desc <- variables[,2]
```

In this sample, we used the library XLConnect to read to content of the Excel document. There is, of course, many libraries to import Microsoft formatted documents but we chose this solution because it's a cross-platform solution that is recommended by the website R-Bloggers. Although they're recommending converting first excel into CSV. (R-Bloggers, 2013)

### 3.5.2.3 Data cleaning

Quality is the major issue in KDD process and this trouble is solved by the cleaning phase. Since the data are not coming directly from a data warehouse but from an accounting system, data are not pre-cleaned. (In DWH process, there is also a data cleaning phase.) That is why in our dataset there is:

- Missing value
- Simplified value
- Use of Different metric system
- Formatting error
- Semantic errors.

All these errors must be corrected if we want to perform a meaningful data analysis on for this company.

#### Missing Values (NA's)

Data mining is really sensitive to missing values. Most of the time, predictive models are ignoring these records and the created models are then incomplete and imprecise. That's why it is really important to clean this data. Missing values, occur when no data value is stored for the variable in an observation (Wikipedia, 2016). By default in R those errors are replaced by NA (meaning not available).

By executing the command **summary()** on our data frame object (summary is used to see information about objects), we can clearly see that there are rows that don't have any values. We can easily delete them from the distribution by executing the following command.

```
### Remove NA
Sales_Master <- Sales_Master[!is.na(Sales_Master$do_numero), ]
```

Sometime systems can also interpret missing values as zero. In our dataset there are lines that have zero quantity and no price. We choose to delete those lines because they aren't relevant for our analysis. The following code is executing this task.

```
#Remove data with missing numerical values
Sales_Master <- Sales_Master[!Sales_Master$d1_qte1 == 0, ]
```

This task was quite easy because we already suppressed most of the NA during the selection phase.

## Redundant observations

The article dataset is containing multiple time the same article because a good sold in 2010 can also be sold during the next year. We have to suppress the old article and only keep the updated one. The code below is executing this task.

```
##get the latest article by update date

#set id to rows
articles$ID<-seq.int(nrow(articles))

#transform update_time to Date
articles$update_time <- as.Date(as.character(articles$update_time),
                                format = '%d.%m.%Y')
articles$creation_date <- as.Date(as.character(articles$creation_date),
                                format = '%d.%m.%Y')
#select NA update_date
na_updateDate <- is.na(articles$update_time)

#replace NA updateDate with their Creation date
articles$update_time[na_updateDate] <-
  articles$creation_date[na_updateDate]

## select the product that has the highest update date
x <- articles %>% group_by(dl_article) %>%
  summarise(id = max(ID), update = max(update_time))
#remove duplicate article
articles <- articles[articles$ID %in% x$id,]
```

First step on that snippet is to create an id for each row on the articles list. Then we need to transform both update and creation date from string to date object. After that we select articles that were never updated and assign them their creation date. Then we use dplyr, a library that allows use to manipulate datasets with a really simple syntax (Rstudio, 2015) to group the same article and get the id of the most recent line. To conclude this step we only have to select the lines with these ids. This sample of code has deleted most of the redundant data, but with a closer look on our dataset we can see that there are still some redundant elements. By using the **summary()** command on the result we can see that it exists some article that has different ID. Below, you will find the result of the summary function.

```
> summary(articles)
  dl_article
Min.   : 3  Champagne Deutz Brut Rosé 75 cl
1st Qu.:2375 Consigne caveau
Median :3798 Port Ellen 32Y 70cl 52.5°
Mean   :3685
3rd Qu.:5186 Brunello Di Montalcino Greppone Mazzi 2007 75 cl
Max.   :6664 Cabernet Sauvignon Les Ferrieres 2011 75 cl
         (Other)

  desc      group
:         :
2         :  V-4004-ROUGE : 553
:         :  SPR - WS - SCT: 326
:         :  VRG - F - BDX : 291
:         :  A-18      : 213
2         :         : 146
:         :  CHP - BL   : 144
:4445      : (Other) :2784
```

Figure 3: Summary on articles

In order to clean those articles without affecting data in the sales document. We must create a mapping table to create a relation between old and new articles. This mapping table will be used to make correspond old id with new id. The following R code is performing this operation.

Code in **prepare\_article.R**:

```
#create a mapping table
x <- articles %>% group_by(desc) %>%
  summarise(id_old = first(dl_article),
            id_new=last(dl_article),
            cnt = n_distinct(dl_article))%>%
  filter(cnt >1)
#save it for further use
write.csv(x = x, file = 'data/article/ID_mapping_table.csv')
#suppres old article
articles <- articles[!(articles$dl_article %in% x$id_old),]
```

Code in **prepare\_sales.R**:

```
#read id mapping table created in article preparation
mapping_table_id <- read.csv(
  file = 'data/article/ID_mapping_table.csv',header = T)
for(i in 1:nrow(mapping_table_id)){
  Sales_Master$dl_article
  [Sales_Master$dl_article == mapping_table_id[i,'id_old']]
  <- mapping_table_id[i,'id_new']
}
```

The code on the prepare articles script is creating the mapping table. To create it, we're taking advantage of the fact that the more recent the article is the higher his id is going to be. This table has the following format: **Description;id\_old;id\_new**. Then we're applying the mapping table on the sales dataset by looping through it. On the loop we are transforming the old article id to new one.

### Semantic errors

Semantic errors are entries that are not following the rules resulting from the variables meaning. These mistakes can easily happen if the database field does not have a constraint that enforces inputs. An example could be a negative value for an age. By looking at our dataset values summary, we can see that some values are not consistent with regards to what they should actually describe.

Below is the summary command that showed inconsistent values

```
> summary(articles[,c('dl_article','size','prct_margin','stock_unit','stock_total')])
  dl_article      size      prct_margin      stock_unit      stock_total
Min.   :    3  Min.   : 0.0000  Min.   :  0.00  Min.   :    0.0  Min.   : -2000.00
1st Qu.:2373 1st Qu.: 0.7000 1st Qu.: 40.53 1st Qu.:    7.2 1st Qu.:    0.00
Median :3796 Median : 0.7500 Median : 43.90 Median :   23.8 Median :   31.23
Mean   :3684 Mean   : 0.9158 Mean   : 44.17 Mean   :  470.9 Mean   :  235.79
3rd Qu.:5189 3rd Qu.: 0.7500 3rd Qu.: 47.73 3rd Qu.:   55.2 3rd Qu.:  239.54
Max.   :6664 Max.   :150.0000 Max.   :200.00 Max.   :1801402.0 Max.   :22888.24
```

Figure 4: Summary of semantic errors

We can rapidly see that size of bottles written in litres because the majority of data, from the first to the third quartile is smaller than 1. The most common bottle size is 0.75 l. It is not likely that the shop owns 150 litre bottles so we can easily conclude that this is a hand-written mistake. To support our theory, we can also represent this variable in a box plot to have a look at the outlier.

Concerning the column `prct_margin`, which means the benefit margin in percent, we can easily see on the summary that something is not normal. To have a better look on these extreme values we can create a **boxplot()** of the variable.

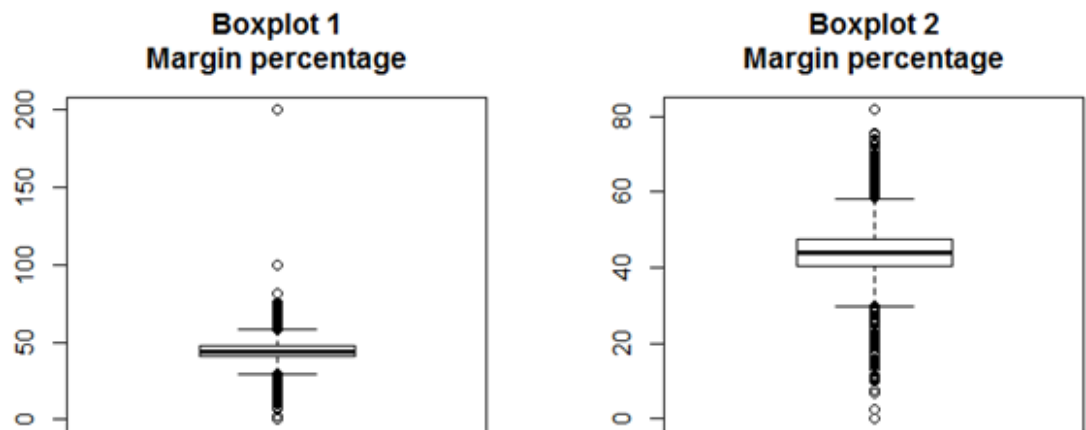


Figure 5 Boxplot of margin percentage with and without extreme values

Here we can observe in the first box plot that the maximum value is really too high compared to the other values. By suppressing this extreme value we have a range that is more acceptable. There is no correspondence to do with the sales document because this article was never used. The above box plot is also showing us the importance of this phase. The same process needs to be done for the other fields.

## Format Data

With the past steps, we have reduced the number of articles to 4447. Our job now is to reformat the data so we can interpret them. In this distribution, there is some interesting attribute such as the alcohol type or the origin country of the product.

```
> summary(articles[,c('group', 'country')])
```

group	country
V-4004-ROUGE : 553	4004 : 731
SPR - WS - SCT: 325	: 552
VRG - F - BDX : 289	18 : 435
A-18 : 212	France - Bordeaux: 271
: 145	15 : 201
CHP - BL : 144	1102 : 165
(Other) :2780	(Other) :2093

Figure 6: Summary of group and country columns

The result of this command showed us that the attribute group and country are not following a precise format. It seems to be a different way of writing the data.

For the group variable, we can see that half of the variable use numbers to represent their second value, while the rest use only alphabetical values. This second variable represents the country because we saw that the 4004 number is used in both country and group variable. Concerning the country column, values seem to be written using both numerical code and country-region name.

When we are selecting only the values that contain a numbered pattern, we realized that those products' update date value is between 2007 and 2013. We can say hypothetically that this is an old format.

```
> summary(articles[grep(pattern = "[0-9]+", x=articles$group),c('group', 'country', "update_time")])
```

group	country	update_time
V-4004-ROUGE:553	4004 :692	Min. :2007-09-06
A-18 :212	18 :212	1st Qu.:2008-08-10
V-4004-BLANC: 92	1102 :134	Median :2010-03-23
A-15 : 89	1104 : 94	Mean :2010-01-11
V-4005-ROUGE: 76	15 : 89	3rd Qu.:2011-04-12
V-1102-ROUGE: 61	4005 : 84	Max. :2013-01-09
(Other) :504	(Other):282	

Figure 7: Summary of the old pattern rows

This information allows us to speculate that the different formats in these fields are related to the time variable. The company have changed their format between 2013 and 2014. This hypothesis can be easily checked by selecting the articles that aren't matching the pattern. We have chosen to convert the old variable into the new format because this format is more self-explaining than the old one.

Now that the rules applied to these formats are known, we must create a mapping table to transform the value of that column. A fully automated solution is not needed because this concerns the old products.

In order to do that, we first have to select and export the different possible values.

This snippet is not in the data preparation script because it is a on time task.

```
articles$group <- gsub(" ", "", articles$group, fixed = F)

#select all group variables
group_variable <- articles %>%
  group_by(group) %>%
  summarise()
write.csv(x = group_variable, file = 'data/article/group_var.csv')
```

After that, we manually recode all the variables in an excel filename variables.xlsx in a standardized way. (Once again, we have chosen to use an Excel document because we want that document to be easily modified. Below is an example of that file.

Source	Target
1101BLANC	VRG-CH-VD
A-2	SPR-EDV-CH
AFRIQUESUDBLC	VL-ALF-SUD
V-1101-ROUGE	VRG-CH-VD

The format is the following: **Alcohol type – Characteristic 1 - Characteristic 2**

We have preferred to classify all strong alcohol together in the group (SPR – Spirit).

It allows us to reduce the number of sub-groups. The type of alcohol for spirit will be saved as first characteristic. Now that the mapping table is created, the following code is applying it to the dataset.

```
#remove space
articles$group <- gsub(" ", "", articles$group, fixed = F)

#loading the excel sheet containing the mappings
mapping_group <- readWorksheet(
  loadWorkbook('data/article/Variables_article.xlsx'), sheet=3)

for(i in 1:nrow(mapping_group)){
  articles$group[articles$group == mapping_group[i,1]]
  <-mapping_group[i,2]
}
```

Firstly, we have to remove the white space on the column. Then we load the excel sheet and apply it to the column using for a loop. Now the whole column is written using a standardized format and we can extract these 3 different values using the function below.

```
#separe group variable using tidyr
articles <- articles %>%
  separate(group, c('alc_type', 'alc_type2', 'region'), '-')
```

To do this, we have used tidyr, it's an R library that is expending the language of dplyr package we have used before (Rstudio, 2014). We have preferred to dplyr over reshape2 because it is easier to use and self-explaining.

The result of the first column can be with the command **table()** used count each combination of a factor level (R Documentation, 2016).

```
table(articles$alc_type)
```

	ACC	BIERE	BONCADEAU	BONCADEAURET	CHP
144	130	23	1	1	250
FOOD	HYDRO	HYDROMEL	SERV	SODA	SPR
45	1	1	16	1	1346
TEST	VBL	VIN	VM	VM	VRG
1	616	3	1	2	1760
VRS					
99					

Figure 8: Table showing the alc\_type values

We can notice that there is types of article that are really not represented such as VIN or HYDROMEL. For analysis purpose we cannot use these variables and therefore choose to create an alcohol type called other that will contain all the unknown and small groups. This R-snippet is regrouping the small size factors and replacing the empty values.

```
## regroup alcohol types that are really small
x <- articles %>% group_by(alc_type) %>%
  summarise(cnt = n()) %>%
  filter(cnt < 10) %>%
  collect %>% .[[1]]

articles$alc_type <- as.character(articles$alc_type)
articles$alc_type[articles$alc_type %in% x |
  articles$alc_type == ''] <- 'OTHERS'

articles$alc_type <- as.factor(articles$alc_type)
```

After these commands, we obtain a consistent article category that can be seen using the **table()** method. The Figure 9 below is showing the output of this phase.

```
> table(articles$alc_type)
```

ACC	BIERE	CHP	FOOD	OTHERS	SERV	SPR	VBL	VRG	VRS
130	23	250	45	162	16	1346	616	1760	99

Figure 9: Table showing the alc\_type final values

We are not regrouping the low represented factors for the two other columns because we consider them as sub-categories. But we are still replacing the empty value with OTHERS. The same process is used to transform both column country and region.

### 3.5.3 Data transformation

Now that the problems concerning data quality are fixed, it is time to recode the variable in a better format. In this section we will see how we have discretized variables, construct new variables and reduce the range of variables.



### 3.5.3.1 Discretize Variables

In our dataset we now have more than 4000 different articles. Each of them has his own price. Some algorithms such as the classification or association rules don't like to work with numerical value. The rpart library from R is automatically converting numerical values into bins and apriori package for association rules mining is only accepting factor columns (CRAN, 2016). By binning articles price, we are also creating a variable that is splitting articles into cheap, expensive and luxurious products.

We chose to separate the variable into five different bins that are containing an equal frequency because our goal here is to bin article into cheap/expensive. Below this R snippet is splitting is used to split this numerical value.

```
#Discretize price
q <- quantile(articles$vnt_price, seq(0,1,by = 0.2))
q[1] <- q[1]-1

articles$pric_Index <- cut(articles$vnt_price, q)
```

Here we are first calculating the different bins by splitting the value every 20%. After that we are subtracting 1 to the first quantile so it is also including the lowest values. Then we are creating a new column in our dataset by cutting the sales prices with the quantile created. The result is the following:

<b>(-1,22]</b>	<b>(22,42]</b>	<b>(42,79]</b>	<b>(79,159]</b>	<b>(159,1.15e+04]</b>
<b>928</b>	<b>857</b>	<b>909</b>	<b>887</b>	<b>866</b>

Figure 10: Result of the price discretization

We now have five different articles group. Each group is a range of price. This discretization will be useful to perform analysis on the cheap or expensive product. The bins name could easily be changed by referencing the name attribute on the cut function. It is also important to notice that we are creating a new column and not overwriting the sales price because we don't want to toss away the prices information.

### 3.5.3.2 Construct new variables

Some variable can be created from initial variables to have a more discriminant predictive model. This is some example of variable creation:

- The age of person can be calculated with his birthdate.
- The birthdate and the date of the first product bought can produce the age of the person when he became a customer

In our case, there is some valuable information that can be derived from the date of sales. With this field, we can create many variables such as the month, year and day of week of the sales. The R snippet below is illustrating this process.

```
# Transform column to date and time
Sales_Master$do_date1<-as.Date(Sales_Master$do_date1, format=
'%d.%m.%Y')
Sales_Master$day_of_week <- as.factor(
  format(
    Sales_Master$do_date1,
    "%A")
)
```

On this code, we are first transforming the string `do_date1` to date object. Then we are extracting the variable day of week by using the `format` function. This function is using the **strptime** format which is, according to R help, a function to convert dates from and to characters (Inside-R, 2016). We are here using `%A` to extract from the date the day of week.

#### 3.5.4 Data mining

Now that we have pre-processed and transformed the data, it is time to find pattern and model on this dataset. This section will present two datamining sub-fields which are the customer behaviour analysis and the time series forecasting. This topic is covering the mining of patterns as well as their evaluation.

#### 3.5.5 Mining Customer Behaviour

It is a great asset for a company to understand the behaviour of their customer. It is helpful for a manager to know what customers like to buy, when and where. It allows him to create offer that is optimized. This section aims to explain market basket analysis as well as a frequent pattern analysis.

## Market Basket Analysis

We will in this section perform a market basket analysis (MBA) on the sales to discover if items are often bought together. The pattern used for this study is the single dimensional association rules because we are trying to find relation between the different items bought. More explanation concerning that pattern can be found on the section 2.5.3 Association Rules.

With this information on the data, Retailer can improve the customer shopping experience by modifying the store layout or create marketing offer to encourage clients to buy more items (Snowplow, 2016). For example, Amazon is widely using affinity analysis to perform cross-selling when they are recommending products based on their purchase history and the history of other people who bought or searched the same product.

For this type of analysis, R is offering us the package `arules` that provide the infrastructure for association rules mining. This package is implementing the Apriori algorithm that is designed to operate only on transactional datasets. Therefore we have to prepare our dataset before using this algorithm.

Our dataset is currently describing every line on the sales receipt. The goal is to transform our data frame into a transaction object containing one line per receipt. This operation is made in two stages. First we have to aggregate our table, and then transform the result into a transaction object.

```
## Aggregate sales receipt
list_Achat <- sales_articles %>%
  group_by(do_numero, desc) %>%
  summarise() %>%
  group_by(do_numero) %>%
  summarise(list = list(desc), cnt = n())

#transform list to transactions
trans <- as(list_Achat$list, "transactions")
```

In this snippet, we are aggregating the rows by grouping them with their receipt id.

The second command is used to transform the list into a transaction object, which is a matrix containing the transaction as rows and items as columns.

By using the command **summary()** on that created object we have an overview of the object. You see below the output of that command.

```
> summary(trans)
transactions as itemMatrix in sparse format with
36236 rows (elements/itemsets/transactions) and
3704 columns (items) and a density of 0.0004244955

most frequent items:
      Champagne Gobillard Brut Tradition 75 cl      1101
      Voucher 15 retour                          743
Aligote Les Perrieres, Peissy AOC Genève 2009 75cl    372
      Champagne R de Ruinart Brut 75 cl      839
      Entrée Soirée Dégustation              399
      (Other)                                53521

element (itemset/transaction) length distribution:
sizes
 1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   20
23806 8013 2678 896 342 231 112  56  30  18  11  11  5  6  7  6  4  2  2

      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      1.000  1.000  1.000  1.572  2.000  20.000

includes extended item information - examples:
      labels
1  champagne Billecart Blanc de Blanc 75 cl
2  champagne Billecart Salmon Rosé 150 cl
3  Champagne Moët & Chandon Ice Impérial 75 cl
>
```

Figure 11: Summary of the transaction object

This output is separated in four sections. The first part is informing us on the number of transactions and objects found on it. Interesting information presented there is the density of the matrix, which is the total number of non-empty cells divided by the total number of cells. The density here is equals to 0.00042 which is really low for a transaction set. That number showed us that our matrix is really empty. The second element of the summary is informing us about the most frequent item on the list. We see that the item ‘Champagne Gobillard Brut Tradition 75 cl’ is the most frequent item. It has a support of 0.3 (1101/36236).

Next information available is the length of the transaction. We see here that more than two thirds of the ticket contains only 1 article. It showed us that customers are buying most of the time only one article. This information is represented with a list containing all the different size as well as a statistical overview with Quartile, median and mean. The fourth part is displaying the first three items in the dataset.

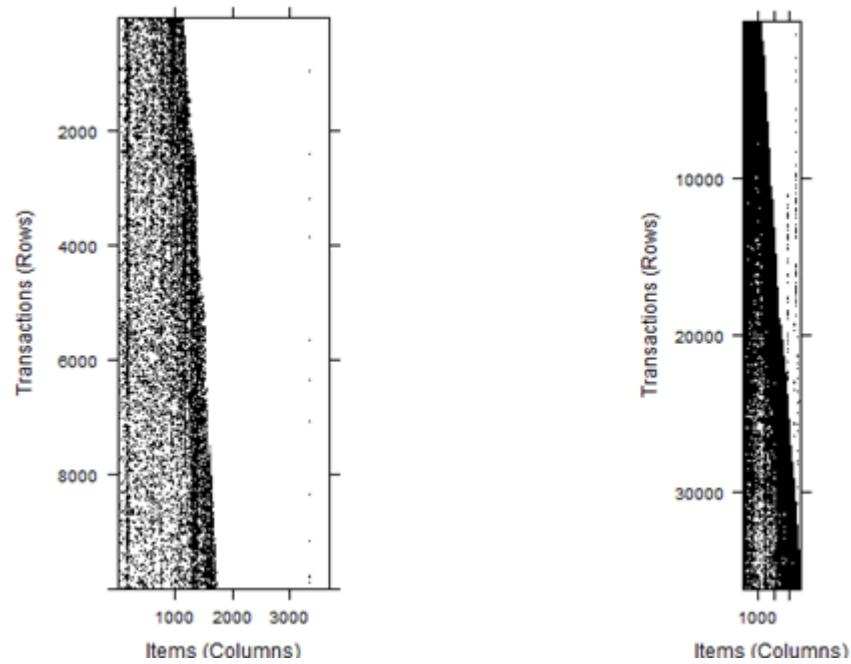


Figure 12: Item repartition on the transaction set for the first 10000 and all the set

By using the command **image()** on our transaction set, we have a direct visualization of the binary incidence matrix where the dark dots represent the item bought on in the matrix. (CRAN, 2016)

According to the left plot (Figure 12 left) that is showing the item repartition for the first 10000 transaction, we see that items in our dataset are not well distributed. The white area on the right side is showing us that more items were added through years. This plot is also informing us that the old products are getting less and less sold. We can conclude that this is not a good idea to process the whole transaction set. Therefore we will limit our algorithm on the 15'000 last transactions.

To mine association rules we have to use the function **apriori** of the package **arules**. The default parameter of that function is to search rules with support 0.1, confidence 0.8, and max length 10. (CRAN, 2016) In other words, rules have to be present 10% of the time and have an 80% confidence to be by default recognized by the algorithm. This is most probably impossible with our dataset because of the low density of our transaction matrix. It is improbable that object can be bought 10 % of the time together knowing that there is more than 3000 articles.

If we execute the function on the dataset, we won't be able to find rules because there are no rules with such criteria. However it is possible to find some rules by adjusting both support and confidence.

The following commands are creating rules using the apriori algorithm.

```
rules <- apriori(
  data = trans[nrow(trans)-15000:nrow(trans)],
  parameter = list(support = 0.0005,
    confidence = 0.1,
    minlen=2))
```

. It is possible with function **summary()** to have an overview of the rules objects, this overview is showing us much information such as the number of rules, length distribution and quality measure of support, confidence, lift. Below the summary on our rules set object.

```
> summary(rules)
set of 135 rules

rule length distribution (lhs + rhs):sizes
  2   3
120  15

      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
2.000   2.000   2.000   2.111   2.000   3.000

summary of quality measures:
      support      confidence      lift
Min.   :0.0005099   Min.   :0.02174   Min.   :  2.317
1st Qu.:0.0005099   1st Qu.:0.09602   1st Qu.: 19.902
Median :0.0006119   Median :0.20833   Median : 41.267
Mean   :0.0007562   Mean   :0.26499   Mean   :106.928
3rd Qu.:0.0008159   3rd Qu.:0.39565   3rd Qu.:112.057
Max.   :0.0026517   Max.   :0.83333   Max.   :1167.262
```

Figure 13: Summary on objects rules

We first see that the apriori algorithm has created us 135 rules, which is not a lot compared to the amount of transaction we have. Secondly, it shows us that most of the rules are binomial. The information concerning the quality measure of the three indicators informs us that the support is really low because the maximum support is 0.02%. Something interesting is appearing. We see that the lift is really high, which means that the rules are really accurate. But it is important to remember that this indicator is calculated using the support of rules as well both item support. It means that a low support can inflate the lift result.

It is also possible to watch the rules on list with the function **inspect()**. We will here look at the top five rows by support to show the 5 most popular item combination.

```
> inspect(sort(rules, by='support')[1:5])
```

	lhs	rhs	support	confidence	lift
99	{Biere 1906 Reserva Especial 33 cl 6,5°}	=> {Biere 1906 Red Vintage La Colorada 33 cl 8°}	0.002651708	0.4062500	73.76447
100	{Biere 1906 Red Vintage La Colorada 33 cl 8°}	=> {Biere 1906 Reserva Especial 33 cl 6,5°}	0.002651708	0.4814815	73.76447
111	{Bière Docteur Gab's Blanche Houleuse 33 cl 5°}	=> {Bière Docteur Gab's Ambrée Chameau 33 cl 7°}	0.001733809	0.2982456	66.46132
112	{Bière Docteur Gab's Ambrée Chameau 33 cl 7°}	=> {Bière Docteur Gab's Blanche Houleuse 33 cl 5°}	0.001733809	0.3863636	66.46132
73	{Gamay De Peissy Perrieres 2013 75 cl}	=> {Aligoté Les Perrières 2013 75 cl}	0.001529832	0.3846154	33.37304

Figure 14: Inspect rules by support

We see that the rules 99 is the following

{Biere 1906 Reserva Especial 33 cl 6,5°}=> {Biere 1906 Red Vintage La Colorada 33 cl 8°}

This is the most popular association on our dataset. It is present 0.2 % time on the dataset. It has a confidence of 0.4, which means that the consequent is bought 40% percent of the time people bought the first beer. It has a lift of 73 that mean that these products are more often found together than separately. By sorting the rules by confidence we can find items that are often bought together.

```
> inspect(sort(rules, by='confidence')[1:5])
```

	lhs	rhs	support	confidence	lift
1	{Biere Darach Mòr Fût Invergordon 25 ans 33 cl 6°}	=> {Biere Darach Mòr Red Hand Sherry Cask 33 cl}	0.0005099439	0.8333333	1167.2619
2	{Biere Darach Mòr Fût Invergordon 25 ans 33 cl 6°}	=> {Biere Darach Mòr Fût Bowmore 17 ans 33 cl 6°}	0.0005099439	0.8333333	817.0833
3	{Bière Pale Ale Brasserie du Virage 33 cl 4.6°}	=> {Bière Blanche Brasserie du Virage 33 cl 5°}	0.0009178990	0.8181818	668.5227
4	{Pichollette De Geneve Ass Blanc 2013 28 cl}	=> {Pichollette De Geneve Ass Rouge 2013 28 cl}	0.0006119327	0.7500000	319.7283
5	{Bière Blanche Brasserie du Virage 33 cl 5°}	=> {Bière Pale Ale Brasserie du Virage 33 cl 4.6°}	0.0009178990	0.7500000	668.5227

Figure 15: Inspect rules by confidence

Here we can see that the rules 1 has a confidence of 0.83, which means that 83 percent of the time people bought the two products together instead of the antecedent only. This information can be used by the manager to make multiple marketing actions such as:

- Place product x next to product y
- Special offer for x and y
- Perform cross-selling

We can see that the appearance of our rules is really low, less than 0.2 % times in the dataset. According to the default parameter of apriori, a rule should be present at least ten percent of the time to be acceptable. Therefore we can say that it is not possible to have representative rules with that dataset. There are two factors that are making this analysis worthless.

The first cause is that there are too many different objects on the dataset. The density of the transaction matrix had shown us that most of the cells are empty. The second argument is the average of items bought by people. As we figure it out earlier with the summary of the transaction, more than 60% of people are buying one object at the time.

As we see with the previous explanations, we can conclude that it is not possible to discover relevant association rules using products this product list. One idea to solve

that problem would be to reduce the amount of variable by grouping them into alcohol category. But it would not give proper rules because in our dataset, people are buying one item at the time. Another idea could be to remove the receipt that contains fewer than 2 articles but the result would not represent the business anymore and would be statistically wrong when these entries represents the majority of the dataset.

That result is understandable because the business reason of the CPJ is to sell middle to luxurious quality products. As explained on the case description, the shop is also located in the middle of Geneva and clients are therefore most probably visiting the shop on foot. These factors are explaining the fact that people are buying only one good at the time. With an additional client dimension, it would be possible to overcome this problem, because we could regroup the sales by customer id and make them use them as transaction.

### Customer habits

Even though, in this case study, we don't have any information concerning the customers, it is still possible to find some interesting information concerning the customer by analysing when and what they are buying. For this analysis, this time we chose to perform a multivariate association rules. It means that, instead of being focus on the basket list, the analysis will be based on multiple variables.

The first step of this search is to find relevant variable that could be usable. For this case we have chosen 2 dimensions, the first is a time dimension to know when the customer are more likely to buy. The variables dimension is the month, day of week and hours.

Concerning the product, the decision is to take alcohol type as variable because it is considerably reducing the number of choices and make the rules more generic. Below you find the dimension with their respective variables.

Dimension	Variable
Time	month, day_of_week, Hour
Product	alc_type

Customer aspect could be added to this study with values such as gender, age groups, localization and origin which would result in rules with finer granularity.



Second step is to prepare the dataset for the apriori algorithm. As explain during the MBA, apriori is only working on transaction. Below the code use for the selection and transformation

```
x<-sales_articles[c('ID', 'Hour', 'day_of_week', 'month', 'alc_type')]
col <- names(x[,2:ncol(x)])
x[, col] <- lapply(x[, col], as.factor)
trans <- as(x[,col], 'transactions')
```

This code will generate on our transaction matrix a column for each different value of the variable. We can illustrate their frequency this with the **itemFrequencyPlot()** function.

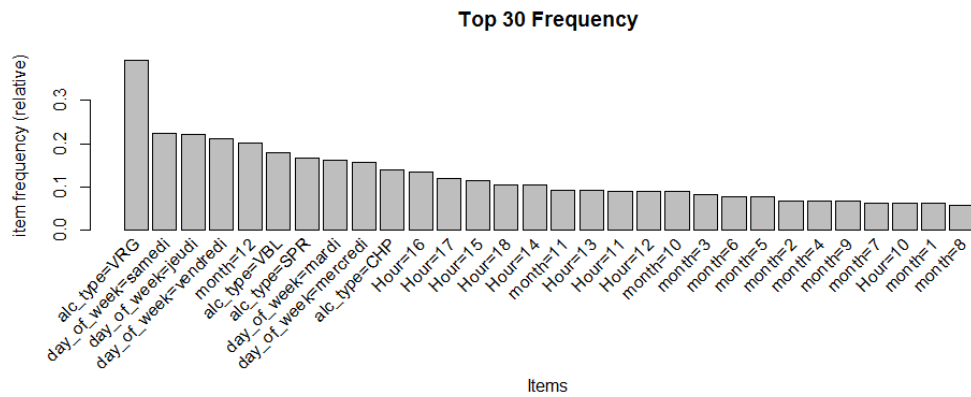


Figure 16: Top 30 frequency of transaction set variables.

On that frequency plot, we can first see that the VRG (Red Wine) is present more than one third of the time, which is really high compared to the other variable. We can already suggest that it would be good to perform an analysis with and without VRG because it would increase the confidence of the other product rules.

Now that we are aware of the frequency of each variable, we can call the command **apriori()** with a minimum support of 1%, confidence of 20 % and a minimum rule length of 2.

```
rules <- apriori(trans,
  parameter = list(minlen=2, supp=0.1, conf=0.2 ))
```

The result of the mining algorithm is a set of 136 rules. The **summary()** function can be used to have an overview of the mined rules. This command showed us the number of rules created, their length, and some statistical information about the support, confidence and lift.

```

> summary(rules)
set of 136 rules

rule length distribution (lhs + rhs):sizes
  2   3
102  34

      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      2.00   2.00   2.00   2.25   2.25   3.00

summary of quality measures:
      support      confidence      lift
Min.   :0.01004   Min.   :0.2004   Min.   :0.8571
1st Qu.:0.01423   1st Qu.:0.2250   1st Qu.:1.0156
Median :0.02030   Median :0.2834   Median :1.0644
Mean   :0.02562   Mean   :0.3235   Mean   :1.2225
3rd Qu.:0.03052   3rd Qu.:0.4012   3rd Qu.:1.1812
Max.   :0.09389   Max.   :0.9968   Max.   :4.9772

mining info:
 data ntransactions support confidence
trans      57100      0.01      0.2

```

Figure 17: Summary of the rules

In this set of rules there are duplicate rules, where x includes y and y include x. It is possible to select only one set of rules by pruning the set. The following code is performing this task.

```

#Remove redundant rules
subset.matrix <- is.subset(rules, rules)
subset.matrix[lower.tri(subset.matrix, diag = T)] <- NA
redundant <- colSums(subset.matrix, na.rm = T) >= 1
rules.pruned <- rules[!redundant]

```

Now that the set only unique rules, we can have a look at the rules created with the function `inspect()`.

```

> inspect( subset( sort(rules.pruned, by='lift'))[1:2])
  lhs      rhs      support  confidence lift
3 {day_of_week=lundi} => {month=12} 0.02639229 0.993408 4.960525
2 {Hour=20}          => {day_of_week=jeudi} 0.02115587 0.989353 4.494197

```

Figure 18: Inspect pruned rules

Here we are facing another problem. The rules are not relevant for us because there is no product involved. Even though we are learning that the shop is open on Monday only in December and that the shop is open on Thursday sometime till eight pm. Hopefully in R it is possible to create a subset of rules that match a pattern. The next figure will show the command and output of that subset.

```

> inspect( subset( sort(rules, by='lift'), subset = rhs %pin% "alc_type=")[1:5])
  lhs      rhs      support  confidence lift
52 {Hour=11}          => {alc_type=VBL} 0.01884413 0.2086080 1.168254
113 {Hour=14,month=12} => {alc_type=VRG} 0.01078809 0.4539425 1.156787
111 {day_of_week=samedi,month=11} => {alc_type=VRG} 0.01078809 0.4444444 1.132583
117 {Hour=18,day_of_week=vendredi} => {alc_type=VRG} 0.01329247 0.4417928 1.125825
40 {month=3}          => {alc_type=VBL} 0.01661996 0.2003801 1.122176

```

Figure 19: Inspect rules with alcohol type as consequent

Here we are creating a subset of rules that have as the right-hand side (rhs) or consequent an alcohol type. The rule 52 is informing us that customers are more likely to buy white wine at eleven am. This is valuable knowledge for a manager because he can use this information to perform cross-selling or advise customers.

As we explained earlier in this section, Red wine (VRG) is too much represented, therefore we chose to exclude Red wine (VRG) from the study because we want to discover relations with other product types. By excluding this variable we discover new relations between variable. This process is creating 184 new rules.

The inspection of the new rules set object give us the following result.

```
> inspect( subset( sort(rules, by='lift'), subset = rhs %pin% "alc_type=")[1:5])
```

	lhs	rhs	support	confidence	lift
156	{Hour=20,day_of_week=jeudi}	=> {alc_type=ACC}	0.01181230	0.4065041	6.879284
5	{Hour=20}	=> {alc_type=ACC}	0.01181230	0.4040404	6.837591
42	{month=7}	=> {alc_type=VRS}	0.01479491	0.2169770	2.593538
54	{month=6}	=> {alc_type=VRS}	0.01621239	0.2022845	2.417917
183	{day_of_week=jeudi,month=12}	=> {alc_type=CHP}	0.01228479	0.3528414	1.513972

Figure 20: Inspect rules with alcohol type as consequent without VRG

We see something really interesting, rose wine is mostly consumed during the months 6 and 7. During the summer, Customers consume more rose (VRS) than usually. This is once again a valuable asset for the company's executives because they know that they should have a sufficient stock for these months.

We can go deeper on this analysis by selecting only on the left-hand side (lhs), which is named antecedent on the theory, the month June and July.

```
> inspect( subset( sort(rules, by='lift'), subset = lhs %pin% "month=6" | lhs %pin% "month=7")[1:6])
```

	lhs	rhs	support	confidence	lift
42	{month=7}	=> {alc_type=VRS}	0.01479491	0.2169770	2.5935382
54	{month=6}	=> {alc_type=VRS}	0.01621239	0.2022845	2.4179168
43	{month=7}	=> {day_of_week=vendredi}	0.01624192	0.2381984	1.1526309
55	{month=6}	=> {day_of_week=vendredi}	0.01768892	0.2207074	1.0679932
58	{month=6}	=> {day_of_week=jeudi}	0.01966748	0.2453943	1.0152456

Figure 21: Inspect rules with June or July as antecedent

Here we see that consumer, during these months are more likely to visit the shop on Friday (vendredi) or Thursday (jeudi). It is possible to plot these rules in another form using the package ArulesViz. Below you see an output of the top 10 rules for June or July

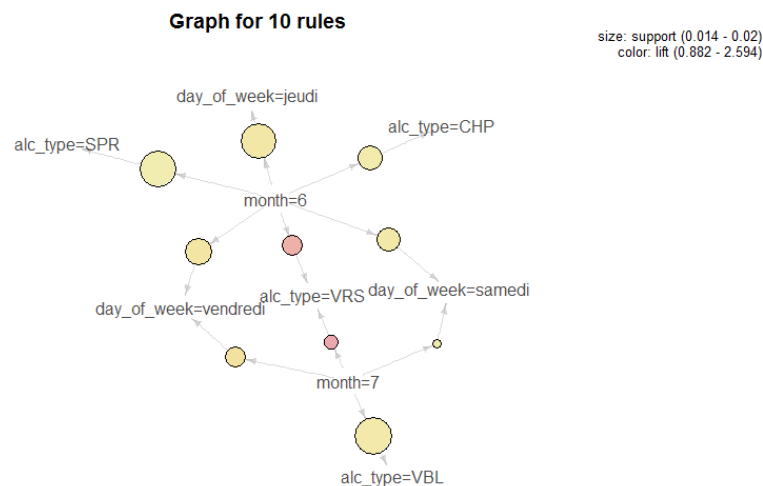


Figure 22: ArulesViz graph of the 10 first rules

We see with the colour that the rules 42 and 54 (Reference Figure 21) have a high lift. The size of the bubble gives us the support, it makes us know that we are selling more SPR, CHP and VBL but this is not the best month for it.

In this section we have seen how powerful this pattern is. We can see the relation between the different values. We learn that we have more chance to sell Rose this time than the rest of the year, which means that if we have some stock at the end of July, we might keep it till next year.

By browsing in the rules created, we can find information to take data-driven decision. Managers could for example organize a summer wine degustation on Friday during these 2 months to make customers buy an additional product.

### 3.5.5.1 Time Series Forecasting

A lot of data on a company are simply an ordered sequence of values and data mining allow statisticians and data scientists to forecast these sequences based on the sequence component. In this section, we will speak about the different algorithm available and aim to forecast the product sales for next year.

This analysis is based on the free e-book “A Little Book of R for Time Series” by Avril Coghlan.

First of all, we have to choose our regular interval. In our dataset it is not possible to perform a daily analysis because the shop is not open every day. A week analysis would also cause some problem because there are many differences between first and the last week of year. That’s why we choose here to perform a monthly dataset.

Now that we know the periodicity, we have prepared our dataset to be able to perform this analysis, this task is done once again using dplyr package. The following snippet is performing the following action.

```
#time series analysis for the number of sales per day
df <- Sales %>%
  group_by(year, month) %>%
  summarise(sum = sum(dl_montant))
```

The next step consist to transform that data frame to a time series object, it is done using the function **ts()**. As we choose a monthly periodicity, we have to set the frequency to 12. We can also specify the first year of our values list. In this case our start date is the first of January 2010. The command below are using the parameter we discuss and the Figure 23 is showing his output. This output can naturally be illustrated using **plot.ts()** like in the Figure 24

```
> tserie <- ts(df$sum, start=c(2010,1), freq = 12 )
> tserie
```

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2010	68829.45	46291.40	75580.55	58191.70	55212.90	62818.35	47785.15	36016.90	58961.25	54219.85	97493.75	235764.75
2011	45927.85	45128.55	53991.70	43090.65	47028.60	57997.10	38351.90	31751.10	49563.30	51729.45	67632.45	206163.05
2012	48957.30	42430.80	55534.85	35660.50	44448.45	47058.80	23859.30	30572.35	40187.15	45205.55	60970.70	155927.25
2013	40301.85	33653.70	41139.10	40212.15	63497.50	57473.55	36451.90	43233.05	39906.50	67005.55	89613.90	200937.60
2014	40319.70	54570.20	54112.30	40526.10	63239.65	56071.60	50341.85	41033.05	48835.65	68247.70	75847.45	181672.10
2015	44637.95	50472.40	55541.20	41688.35	59997.40	65728.25	38415.10	43387.40	56517.30	65478.90	71074.60	170113.45

Figure 23: Time series output

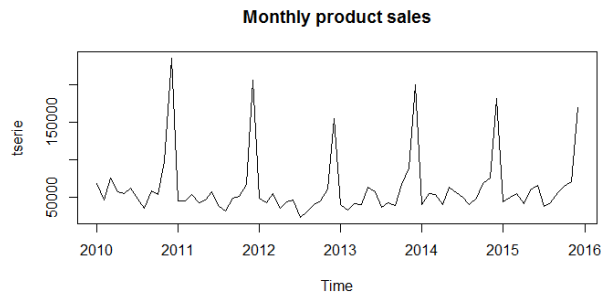


Figure 24: Product sales time series

We see from this time series that there is a high seasonal variation. The amount sales highly depend of the period. During Christmas, we can observe a lot of sales compared to the middle of the year. We can see that these variations are really constant and that there seems to be no long-term trends on that graphic. In that case, there is no need to transform the dataset using logarithm or other mathematical technic because the fluctuations are not increasing over the year.

As our dataset is seasonal, we have other tools to visualize our distribution. As explain in the first part of this thesis (see 2.5.5 Times Series), a seasonal time series is described using trends, season and random fluctuations. R provides us the function **decompose()** that estimate these components. On the figure below is representing the decomposed time series.

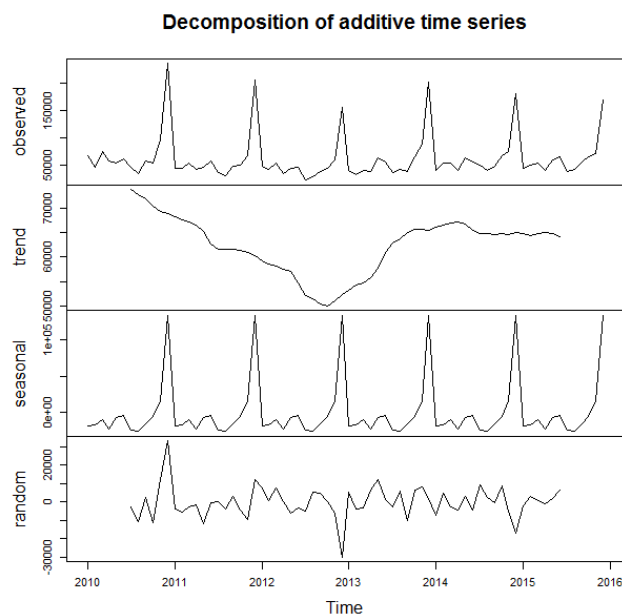


Figure 25: Times series components

The first part of this plot is showing us the original time series, then the trend, seasonality and random fluctuations. The trend is showing us that between the years 2010 and 2013 the amount sold decrease from 70000 to 50000 then went up again till 2014 where it starts to fall again. What we can conclude from this plot is that values appear to be constant over time. Therefore we assume that this set of values is stationary. By definition a stationary time series is one whose statistical component is all constant over time (Duke University, 2016). It is quite easy to deal with because we predict that the component will remain the same as they had been before.

The next step of the research is to select an algorithm that suits this case. To fit the analysis needs, we choose the seasonal autoregressive integrated moving average which is most commonly called seasonal ARIMA. It is possible to modify a non-stationary time series to make it acceptable for ARIMA. It is called differencing.

Below a description of that model proposed by the website forecasting solutions:

“ARIMA methodology also allows models to be built that incorporate both autoregressive and moving average parameters together. These models are often referred to as "mixed models". Although this makes for a more complicated forecasting tool, the structure may indeed simulate the series better and produce a more accurate forecast. Pure models imply that the structure consists only of AR or MA parameters - not both.

The models developed by this approach are usually called ARIMA models because they use a combination of autoregressive (AR), integration (I) - referring to the reverse process of differencing to produce the forecast, and moving average (MA) operations. An ARIMA model is usually stated as  $ARIMA(p,d,q)$ . This represents the order of the autoregressive components ( $p$ ), the number of differencing operators ( $d$ ), and the highest order of the moving average term. For example,  $ARIMA(2,1,1)$  means that you have a second order autoregressive model with a first order moving average component whose series has been differences once to induce stationarity.” (Forecasting solutions, 2016)

As explain on this definition, ARIMA is composed of 3 components, Autoregressive components (AR), differences and Moving Average (MA). Seasonal version of ARIMA looks like the following  $ARIMA(p,d,g)(P,D,Q)_m$  where the first couple is describing the non-seasonal part the second couple the seasonal and the  $m$ , the periodicity which is in our case 12.

These parameters can be found using the **auto.arima()** method created by Rob J. Hyndman. They can also be found using plot called correlogram and partial correlogram. Here we will take the simplest path and run **auto.arima** on our transaction set and save it on a new object

```
> fit <- auto.arima(tserie, trace = T)

ARIMA(2,1,2) (1,1,1) [12] : Inf
ARIMA(0,1,0) (0,1,0) [12] : 1307.36
ARIMA(1,1,0) (1,1,0) [12] : 1296.015
ARIMA(0,1,1) (0,1,1) [12] : 1286.186
ARIMA(0,1,1) (1,1,1) [12] : 1288.489
ARIMA(0,1,1) (0,1,0) [12] : 1289.776
ARIMA(0,1,1) (0,1,2) [12] : 1288.484
ARIMA(0,1,1) (1,1,2) [12] : Inf
ARIMA(1,1,1) (0,1,1) [12] : 1288.284
ARIMA(0,1,0) (0,1,1) [12] : 1303.504
ARIMA(0,1,2) (0,1,1) [12] : 1288.25
ARIMA(1,1,2) (0,1,1) [12] : Inf

Best model: ARIMA(0,1,1) (0,1,1) [12]
```

Figure 26: auto.arima trace

In this screenshot we see that this command have tested all possibility and choose the one with the lowest AICc (Akaike information criteria). This criterion is estimating the quality of a model by estimating the information lost. (Hu, 2007)

Now that our model is created, we can use the **forecast()** method to predict the next 12 months.

```
> fc <- forecast(fit, h = 12)
> fc
```

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Jan 2016	38456.48	22875.28	54037.67	14627.097	62285.86
Feb 2016	44497.28	28146.33	60848.23	19490.662	69503.90
Mar 2016	48949.45	31863.39	66035.51	22818.572	75080.33
Apr 2016	36108.51	18317.69	53899.34	8899.796	63317.23
May 2016	55252.39	36783.68	73721.10	27006.929	83497.85
Jun 2016	56910.36	37787.77	76032.95	27664.887	86155.83
Jul 2016	35512.26	15757.43	55267.09	5299.857	65724.66
Aug 2016	37103.10	16735.64	57470.55	5953.763	68252.43
Sep 2016	47550.59	26588.40	68512.77	15491.693	79609.48
Oct 2016	60243.89	38703.40	81784.39	27300.546	93187.24
Nov 2016	68502.87	46399.18	90606.56	34698.199	102307.54
Dec 2016	170826.02	148173.14	193478.89	136181.428	205470.60

Figure 27: Forecast output

On the upper figure, you can see the output of the forecast method with ARIMA. It is creating us a forecast for the next 12 months, as well as predictions intervals for these values. By default ARIMA is creating predictions intervals for 80 and 95 percent. We can have a better look at our predictions by plotting it.

```
#plot it
plot(fc)
+lines(tserie,lwd=2,col="red")
legend(2014,250000,
      c('Forecast','Real values'),
      lty=c(1,1),lwd=c(2.5,2.5),col=c('blue','red'))
```

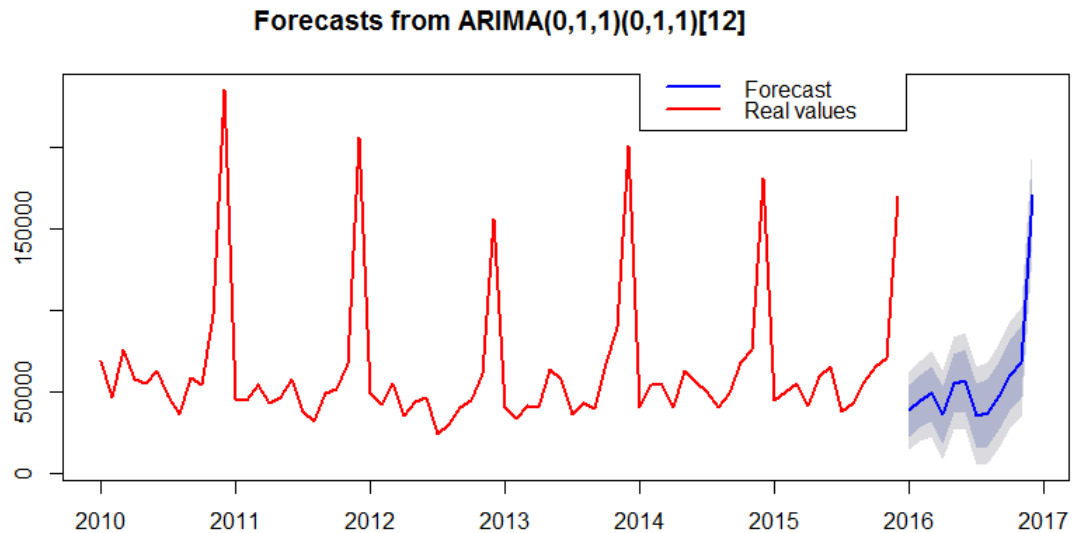


Figure 28 Sales forecast using ARIMA

Our Forecast seems to follow the previous year quite well. We see that the product sales are going to be a bit lower this year. This is due to the trend that was starting to decrease. The information created with this analysis is important for a business because it helps them to anticipate next years. It makes them understand that they are going to have less revenue for the next year. Managers can then take a decision to take a special employee for the end of the year or think the opposite way and fired someone for the same benefit. He can also see that there's an action that needs to be done to revitalize the company and make it grow again. We can also make this forecast for a special wine type to know which product the company should invest on. But it is important to remember that we will have to redo an analysis again and in the case change for a more suitable algorithm.

### 3.5.6 Knowledge Presentation

This final task aims to present the information we created from the whole process. In this section we will explain the software we used to represent this project using dashboard as well as the output of this phase.



According to the case objective, the manager of CPJ wanted to have data visualization to represent his company, he wanted us to create a dynamic dashboard to present his data.

During the KDD process, we have created many graphics to represent this dataset. These graphics were created using the ggplot2 library and the libraries specified plot such as ArulesViz and FrequencyPlot. Ggplot2 is the most famous library for data exploration. It permits us to create a lot of different appealing graphics.

These graphics give the manager a good insight concerning the state of the company, but a problem is that these visualizations are static. It is not possible for the manager to use filters and drill-down options because these graphics are saved as images.

As we were running out of time concerning that project, Dynaxis decided, with the agreement of CPJ, to use Tableau Software to create a dashboard. This option is really interesting for the consulting company because it also gives them a benchmark of that tool. Dynaxis is not currently using Tableau software, but they are hardly considering it, knowing that this software is for 4 years the leader on the market according to Gartner (Tableau Software, 2016).

Tableau Software gives us a user-friendly environment to create data visualization. This software is offering five different versions of their tools. Two of them are the desktop version and the Reader version. The first version allows an end user to create the dashboard. It has an evaluation period of 15 days. The second version, on the other hand, is free and allows the user to visualize Tableau dashboard.

We decided to use the evaluation version to create the dashboard and let the end user browse it using tableau readers. That way, the manager will not have to pay for software and have professional visualization.

We choose to create the following dashboard for our dataset.

- Types and Origin of products
- Sales forecast
- Analysis of the margin per product, origin and price group
- Margins per Hours and days of week.

In this chapter we are going to present only the first two dashboards. This screenshot as well as the other dashboard can be found with a larger scale on the attachment section.

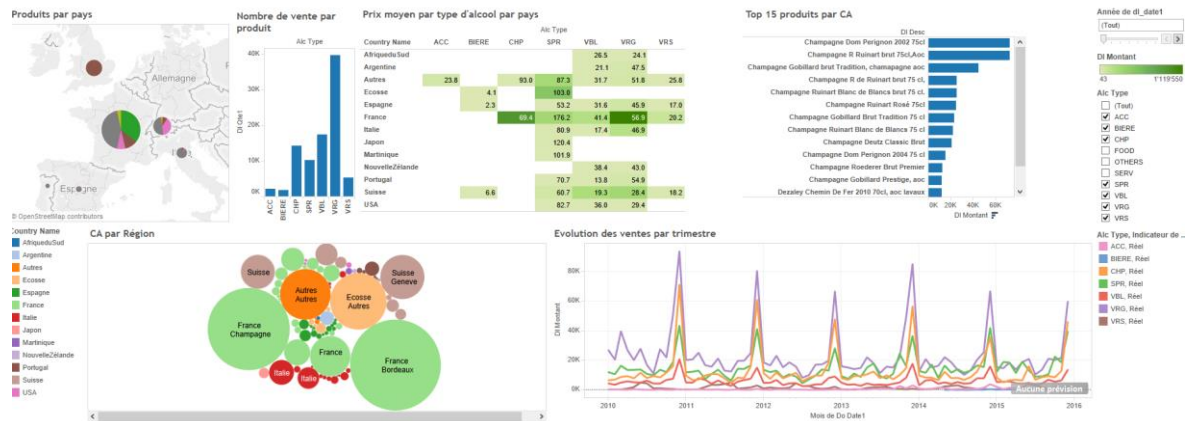


Figure 29: Dashboard of products, origin and type

The first dashboard is separated in six views. The view on the top left corner is showing the product origin on the map. It is giving us information about the product type, amount and origin using a map. Next to it, the second graph is presenting the frequency of the different wine type. Then we have the average price by country and wine type and the top 15 products.

The bubble chart is showing us the power of different regions by country. And the last chart is showing us the trend per product type. On the right we have also some filter that allows the manager to look for a specific year or product type. He can also perform drill down by clicking on all graphics. So he can see for example the evolution of red wine in Switzerland.

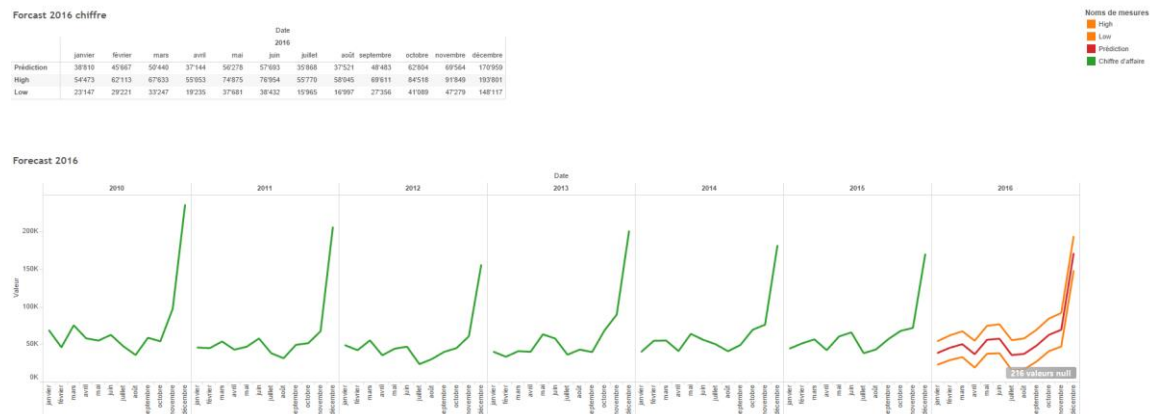


Figure 30: ARIMA forecast on Tableau

The other dashboard we want to present is the sales forecast. Here we are presenting the data calculated during the time series analysis. This dashboard is presenting first the data calculated in a table which includes the forecast value as well as the high and low bounds. Under it there is the forecast representations showing in green the actual values, in red the forecast and orange for the boundaries.

## 4 Discussion

### 4.1 Result of the case study

In this section we are going to explain the result of the case study, we are going to present the result in three parts because every project entity has some result and lessons learned.

We delivered to the customer three different output from this case study. First of all, we delivered him an R project. With it, he can process his data automatically and have an insight of his sales forecast and association rules. Even though the result of the market basket analysis was not as conclusive as we expected, he can still find some relation between his products.

Secondly, we give him a professional dashboard that can be used to have a closer look at his company. CPJ was really happy because they were not able to produce such graphics with their data. It makes them realize the importance of cleaning data and have a better description on their database object. They see that they were able to analyse their sales by alcohol type, region or country. This motivated them to take a step further in this direction. We were expecting that behaviour, so during the final project presentation we have also presented them our suggestion. Those can be found on the Problems encountered sections. In it, we propose them a new way to arrange the data information within the database, as well as technic to collect customer data. The manager was happy because it gave him new objectives for his company.

Concerning Dynaxis, they have learned that predictive analytics is a difficult field, that require a lot of statistical and mathematical knowledges. They have also understood the data mining process, the different pattern that can be found on a dataset and what open-source software are capable of.

As R is really used in the other data mining software as scripting language, it was positive for them to have a case study that presents them some package they can use.

With this case study, they have an example of data mining project which can be useful to calculate how many time they will need for a further project.

They have also a benchmark on Tableau Software, which was really valuable for them because they are considering using it. From this benchmark, Dynaxis has also analysed the potential of their tools compared to that concurrent software. They saw what was better in Microstrategy compared to Tableau and oppositely.

Finally, this case study was a good insight on the data mining field. It makes us first realize the time we need to plan per phase for a similar project. We gain the knowledge of the difficulty that can arise from a small business project. We also learned that the time for the pre-processing phase should be way higher than in a normal project because there is no previous processing.

It gives us a vision on R as data mining software as well as Tableau software. Thanks to that another project would be way faster because we already know which library to use as well as which step we need to make.

## **4.2 Problems Encountered**

The result of the case study was not as promising as what we expected. We encountered a few problems that will be detailed on this section as well as the solutions we suggest to overcome them.

### **4.2.1 Data Quality**

Data quality is really an important aspect of data mining. According to the author of “Data mining and decisional statistics” Stéphane Tufféry, data pre-processing is representing twenty percent of the project. It is a primordial phase because as we explained on the data pre-processing section, it is not possible to produce good statistics with missing information (Tuffery, 2015).

In this business case, there were a lot of issues with that point because the dataset was a direct extract of the accounting system. Below we will present all aspects of that trouble and present solutions for each of them.

#### **Hand-writing**

On this database, the first problem we encountered is that all data were handwritten without any boundary for the user. It is then possible by mistake to enter many wrong values. It is even more likely to happen when the accountant has to enter a long series of products. Samples of these errors can be found on the pre-processing phase of the case study.

Errors such as the bottle size could be easily avoided by setting constraints on the database columns. For that case a constraint that enforces the user to set this variable with a decimal value between 0 and 20. It should also be precise on the field that it

represents quantity in litres. Country and alcohol type could also be bounded with an enumeration that would only allow certain input. In the accounting system, those solutions would be implemented using drop down lists.

### **Missing Values**

If we can process hand-written data to convert them to acceptable values, there is nothing we can do with missing values. In important fields such as country and region, more than 20 percent of the values were missing (20% for country and 36% for the region). It renders these fields difficult to use for analysis. These variables were important in order to classify the different object and with the loss of such an amount of values, we can no longer use them for data mining purpose.

To ensure that the accountant filled these fields, we do not have to create null constraints on the database and brief the team about the importance of data quality for further use. The manager could also create reporting with all the incorrect fields and ask the service responsible for it to correct it.

### **Item Description**

Another problem on this dataset is that there is too much information stored on the articles description. Below a product description that is explaining the problem.

#### **Château Angelus 2006 St Emilion AOC 75cl**

We have for this description 5 different information. Château Angelus is the company that produce that wine, 2006 is the vintage, St Emilion is the name of the wine, AOC is the official label and 75cl is the bottle size. This is really difficult to automatically separate these fields because another product can have another amount of information and those cannot be split by a comma separator.

Here we are losing a bunch of information, we cannot analyse only the wine produce by a cellar or only the AOC labelled wine. Another concern is the database performance because we save two time vintage and size of the bottle. Therefore we suggest for the company to split this information in five fields. As the vintage, product description and size already exist, they should only create new columns for the cellar and label. This would allow them to have a better insight on their product.

The other problem we explain during the pre-processing phase of the case study is the item group. We suggest they have a clear group nomenclature like the following.

- Product type with 8 level like we created in the dataset with SPR, VRG and VBL.
- Product type 2 that should be used with spirit to inform about the type of spirits such as VDK for vodka and WS for whiskey.
- Countries should be written using the ISO 3166-1 alpha-3 code such as FIN for Finland.
- Region should be written using their corresponding code like VS for Wallis or ZH for Zürich.

All these suggestions would largely help CPJ to have a better insight of their sales and products. It would also be a great asset for further statistics and data mining project. By using such item group, we would also limit the handwriting error.

#### **4.2.2 Lack of Dimension**

As our customer is a small company, they are not storing a lot of data. On their database, CPJ is only storing the articles and both sales and purchase related to it. As explain on the case study description, they are not currently storing any information related to the customer. That kind of data would be really interesting for them in many aspects. First of all, they could create a mailing list to invite their customer to some event the company is organizing. With the help of data mining they could then know which customer is more likely to come and focus their marketing operations.

This information collection could be easily made by customer loyalty program. A CLP, according to Investopedia, is a very basic marketing idea that aims to reward customers who are frequently making purchases (Investopedia, 2016). The goal of this method is to motivate the customers to buy more by giving them reward such as deductions or free items. More than two thirds of the small business that is using this technic are saying that this makes more money than it cost to maintain it (Entrepreneur, 2014). Beside the financial part, company can also collect a lot of data such as names, email... But most important, it would allow the company to group the sales by customers, and have better output concerning the market basket analysis. Therefore, we are warmly suggesting our customer to create a fidelity card because it would help them to collect data and enlarge their sales.

### 4.2.3 Communication

Communication is the biggest non-technical issue we endure during this project. One third of the time, poor communication lead to a project failure and it has, according to Coreworx more than half of the time a negative impact on a project. (Coreworx, 2013)

In this project we had that kind of problem because we were not able to have a direct communication with both customer and database administrator. As explained in the project description, the database administrator (DBA) which is a third party didn't want us to have a direct access to the database. He has preferred to give us excel extract of the database. This person was most probably not informed on the project goal and the role he was playing on the project. Therefore his interest was not sufficient enough to help us conduct this project. It was really difficult to speak with him, so we had to wait sometime many days to receive the extract we needed. The direction of Dynaxis, did not want us to pressure this third party because it was a pilot project and as a consulting company, we need to be patient maintain a good relation with the client. This cause some delays to the project because we were waiting for him to receive datasets.

A second aspect of that problem is the lack of communication with the CPJ worker. As stated in the Cross Industry Standard Process for Data Mining (CRISP-DM) process, the first phase of the process aim to understand the business activity, the context and situation of the company (Brown, 2014). On that project we have performed this step during a face to face meeting with the manager but he would have been good to have the point of view of the worker, because they could give us another point of view concerning that company.

Consequently, we propose for the next project to have a direct access to the database, and that the DBA should work as an advisor for the consulting company. Secondly, all parties involved should be represented during the business understanding part because they could give their input and the brainstorming would be more productive. It would also permit us to explain them the project, what we are looking for as well as the advantage they can earn from it. With a face to face meeting, we could significantly increase the interest of all parties involved in the project and they would know what we are expecting from them.

#### 4.2.4 Time

Time is one of the common problem during a project. As this project was a pilot, we had no clue how many time we should allocated to it. We have then fix the project time to one and half month. We have then separated the time according to the literature. Therefore 20 percent of this time was plan for the data pre-processing. But as we started the project, we rapidly see that this was not enough. This phase took 40 percent of the time because we had to learn all the R function as well as data mining theory. The amount of work was also higher due to data quality problem. When we were finished with that step, we had to reduce a bit the data mining part and we had to drastically cut in the data visualization part.

As a result we did not have enough time to create an open-source data visualization. It was plan to create either a dashboard with R package plotly or to create JavaScript plot using the well-known D3.js library. As this project was limited by the internship time, we were not able to extend it. Therefore we choose Tableau Software to produce a dashboard rapidly.

For another project, we advise allocating more time when it has to use open source software, because there is more configuration needed than other software. It is also important to have an audit on the state of the database before planning the project because the phase length is really depending on it.

It is also true that another project would be done faster because we have a lot of lessons learn. We don't need to read all the documentation and literature to choose a package or learn the process.

### 4.3 Is datamining available for small companies?

As we saw in this document, data mining is an important analysis process to explore data. This term as well as big data is currently making a huge buzz on multinational companies. Big firms like Amazon have gained an incredible competitive advantage over their concurrent with these technologies. But these gigantic companies aren't representing the market. According to the Swiss confederation, there was in 2008 only 2.4 % of the company that had more than 50 employees. Our question here is to know if this statistical field is also available for the rest of the market. To answer this question, we will base or reflection on sale companies.

Today, Excel is the most used software for analysis because it is the easiest way to transform data. According to the website wpcurve, in 2012, only 16 percent of the



small company and 33 for the medium were having a business intelligence infrastructure (Wpcurve, 2012). In 2015, both big data and Business intelligence have made their way to the top 10 IT priorities of company (Techaisle, 2015). It proves us that small and medium business (SMB) are now willing to invest on data management. We can conclude with this number that this sector is really evolving and transforming itself.

Business intelligence levels are regrouped in a maturity model that show how develop the infrastructure is. Below we have an insight of the BI maturity model.

0	Limited BI / Spreadsheet
1	Operational Reporting
2	Query & Analysis
3	Dashboard Management
4	OLAP
5	Data mining

According to the findings we made we can say that now more and more company are aiming to have a BI solutions, which is in this maturity model represented by the third or fourth level.

As we see, data mining is the very last step of that process. Before it, on the fourth level, we have the OLAP level with stand for online analytical processing. In field the goal is to create data cubes in order to improve the query respond time. This maturity model makes us quickly understand that it is really difficult for a company to set up an automated data mining layer without having reached the previous level. Therefore companies should first have a system that is collecting and ordering the data before starting to think about data mining analytics.

Now that we know what does the market want, and what is technically needed for a data mining research, it is time to see what are the available data for these small companies.

First of all, SMB have business-related data stored in an accounting system. In Switzerland all companies that have a turnover higher than 500'000 are forced to have a formal accounting (Art. 957, Code des obligations, 2016). This accounting system is recording the sales and purchase as well as article list.

Secondly, most of the time, they have a list of clients, coming from customer loyalty program, that can be stored whether on a CRM program or directly in the accounting

system. Then, they have access to data produced by their websites in form of log files and finally they have data coming from social media.

This showed us that even small companies have access to a wide range of dimensions to analyse. With all these points we can say that data mining is available for SMB that match at least these criteria.

With BI infrastructure company can gain knowledge and what happened and why. They can keep track of their business health, sales and purchases. But that technology is not giving them a deeper understanding of the factor that makes, for example, one customer churn or which cross-selling option is more likely to work with a customer. This information helps them to have a competitive edge over their concurrent.

Data mining software producers have understood that small businesses are more and more willing to have insight of their data as well as explaining these results.

Therefore they are aiming to make tools that are useable for SMB companies.

Google for example, with their free website analytic platform called Google Analytics, let the user use big data solutions to understand website visitors' behaviour.

These tools help companies to analyse traffic on their website to make data-driven decision (Buisness news daily, 2014). Other tools such as SPSS, SAS or open source software are trying to democratize the use of data mining.

But Data mining is an investment for these companies and they might do not clearly see the advantage that those kinds of research can provide them.

As we see in the case study, it is also possible to do data mining research without Business intelligence. But company to be interested in data mining need to have, in our opinion, already access to dashboard and reporting capabilities.

With the evolution of IT and data collection, small companies are now capable of doing data mining, but they need to have already a strong IT infrastructure to be able to take fully advantage of it. We can be certain that data mining is available for small firms and his democratization for SMB will keep increasing but it is first the company choice to invest on these technologies. This field is already ready to be used by small companies but the company market is in our point of view not mature enough to use data mining. When the market will be used to BI, it will be able to upgrade to data mining.

#### 4.4 Is open source data mining a viable choice?

In every IT domain there is a possibility to use open source software. In some cases, it appears that the open source version is much more powerful than the proprietary version (Apache2 vs IIS). But most of the time they appear to be more difficult to use than a licenced option. In this section we will measure how effective open source is compared to other software and if they are viable for a professional use.

To answer this question, we will have a look at the data mining market itself. We will base our market analysis on the Gartner Magic Quadrant (MQ). It is the name of a series of market research which goal is to position software on a market against their competitor. This graphic is separating the providers in 4 categories that are the challengers, leaders, visionaries and niche players (Gartner, 2016). We have below the 2016 MQ for Advanced Analytics Platforms.



Figure 31: Gartner 2016 MQ for Advanced Analytics Platforms (KDnuggets, 2016)

As we see in this graphic, on the leaders' side we find some well-known software such as SAS, Dell and IBM SPSS. But the important point here is that we are also findings some open source software like KNIME and RapidMiner. It shows us that

open source software is not just present on the market but they are also really well graded. All these software are integrating open source programming language like Python or R.

As we see, there is good open source software in that field but are they easily usable? All of them are using a node by node workflow management which make it easier to use. There is no need to code a lot to do a project work. These software are really user friendly and allows users to use pre-created tasks such as association rules learners or transpose a table. There is, of course, some other open source software that can be used such as R or Weka for data mining. Those are much more difficult to use but there is much more information available on the web to face problems.

Open source software is always an option that need to be considered when doing a project. Concerning open source software in data mining, there is, as we see on MQ really good and competitive tools available. For someone that is looking for ease of use rather than complete documentation we would highly suggest they take KNIME or RapidMiner. On the other hand, if there is need to understand every step and output of method, it is better to use R because there is a much bigger community for it. A last point is that this is possible to use both because nearly every Analytics software is integrating R as a scripting language.

## 4.5 Evaluation of own-learning

Data mining is a really complex topic that mix multiple knowledge such as databases, machine learning or statistics. To understand it, you have to be aware of all these fields. As my company was not knowing this field, I have to first read some book to understand this field. I have used the following book for my own-learning.

Data mining et statistique décisionnelle - ISBN : 9782710810179

Modélisation Predictive et Apprentissage Statistique avec R - ISBN : 9782710811589

Statistiques avec R – ISBN : 9782753519923

These books gave me a good understanding of the data mining field, scripting language and project management.

After that with R, I based my research a lot on the R-bloggers website or directly on the language documentation. Youtube videos and stacks exchange post also helped me in a lot for both technical and theoretical understanding.

The most difficult part of this learning process was to understand the statistical aspect behind each function. Every output is full of calculation and measurement. I am not convinced that all the decisions I made were correct because it was my first project in this field. As I spoke with some data analyst, they told me that this field require at least a master or PhD in Mathematics and statistics. I think that I learned a lot from that field including R language but this topic require more than 3-month internship to master it.

For the people that are planning to learn both data mining or R programming, I am strongly suggesting you take the datacamp course, watch video and YouTube and have a professional book in order to see all points of view. Data camp is a website that is giving the best tutorial on R language and data analysis. It cost around ten to twenty euro per month. (DataCamp, 2016)

## 4.6 Methodology

From the topic of this thesis a lot of knowledge needs to be assimilated in order to fully answer this research question. In order to answer them and conduct a data mining project, we first had to have a good understanding of data mining. The books presented in the section 4.6 Evaluation of own-learning permitted us to have a good first look on the subject. In these literature, the importance of R in data mining was rapidly getting bigger and bigger. As these books were explaining their point of view, we have also searched some other sources to cross the information as well as learn R language.

To learn that programming language, website like Code School are providing gentle introduction to the subject. Other applications, such as Kaggle, are providing dataset and script to learn data science. We have followed some research and visualization on that platform, then we start to practice R Programming.

Further R learning was done with the help and documentation of that language because it is really well-structure and easily accessible with the command **?Function-Name**. If the answer cannot be found directly on the help, other sources such as Stack overflow are providing solutions for almost every problem.

After that, we looked at the different software available. Two of them (KNIME and RapidMiner) were tested but the community was not big enough to answer the questions we were having.

As we see in this section, we have chosen the most difficult software so we have easily access to resources and help. This clearly shows us that the methodology used in this thesis is a constructivist method. Knowledge on this topic was created by experiences and real case study rather than learning directly from a teacher. This way of working is really different from the school way of learning because critical thinking and independent learning are key factors. If a constructivist method has allowed us to gather a mass of knowledge, we think that an expert on the subject could be helpful to facilitate some theoretical point, especially in the statistical domain.

## 5 Conclusion

Main goals of this research was to prove that small business can also perform analysis on their data, that today's software are available for small business in form of open-source software. In this thesis we first have had an insight of the theoretical part, what can be used and what can be found. It has shown us that this gigantic amount of data will keep increasing and that data mining today is an unavoidable way to keep an eye on data.

Then with the case study we saw, with a simple accounting database, that it was even possible to perform analysis on it. We saw all steps of the KDD process with explanations. As the result was not as promising as we expected, we discovered many problems that can arise from a project. These encountered complications are important lessons learned and a next project should really benefit from them.

Then we discussed the question of the data mining availability for small companies. We saw that the market was willing and transforming itself to a data-driven market. As the today's goal for SMB is the Business intelligence, we can expect them to be interested in data mining as soon as they are going to reach this level.

As with saw on the Gartner Magic Quadrant open source software is now more and more leading the market. With their community commitment, they will evolve faster than other licenced software. This shows us that the software creators are ready to democratize this field for all kinds of business, small or big.

If data mining is not today in all small business priority, it will certainly come when they will be mature enough. At some point this technology will be a must have and companies who won't have it will struggle to survive. We want to conclude this thesis by this quote: "Intelligence is based on how efficient a species became at doing the things they need to survive." (Darwin, 1859)

## 6 Bibliography

- SAP Predictive Analytics. (2016, 01 19). Automated Analytics. *Help*. SAP.
- Art. 957, Code des obligations. (2016). *Code des obligations*. Retrieved from admin:  
<https://www.admin.ch/opc/fr/federal-gazette/2012/59.pdf>
- Badie, F. (2016, 03 18). *Data Integration and Data Transformation (in KDD)*. Retrieved 03 18, 2016, from SlideShare: <http://fr.slideshare.net/farshadbadi/data-integration-and-data-transformation>
- Betegy. (2016). *Front page*. Retrieved from Betegy: <https://betegy.com/>
- Britannica. (2016, 03 11). *Britannica*. Retrieved 03 11, 2016, from Data mining:  
<https://www.britannica.com/technology/data-mining>
- Brown, M. S. (2014). *Data Mining For Dummies*. For Dummies.
- Business news daily. (2014). *6 Big Data Solutions for Small Businesses*. Retrieved from businessnewsdaily: <http://www.businessnewsdaily.com/6358-big-data-solutions.html>
- Business Insider. (2015, 08 19). *Here's how much data is created on the web every minute*. Retrieved 03 11, 2016, from Business Insider:  
<http://www.businessinsider.com.au/infographic-heres-how-much-data-is-created-on-the-web-every-minute-2015-8>
- Coghlan, A. (2015). *A Little Book of R For Time Series*.
- Complexity Media. (2015, 09 11). *DIKW Pyramid Explained*. Retrieved 03 21, 2016, from YouTube: <https://www.youtube.com/watch?v=6VmxaaVwmdI>
- Complexity Media. (2015). *DIKW Pyramid Explained*. Retrieved from Youtube:  
<https://youtu.be/6VmxaaVwmdI>
- Coreworx. (2013). *PMI Study Reveals Poor Communication Leads to Project Failure One Third of the Time*. Retrieved from Coreworx: <http://www.coreworx.com/pmi-study-reveals-poor-communication-leads-to-project-failure-one-third-of-the-time/>
- CRAN. (2016). *A computational environment for mining association rules and frequent item sets*. Wien.
- CRAN. (2016, 03 19). *Package 'arules'*. Retrieved 04 04, 2016, from CRAN:  
<https://cran.r-project.org/web/packages/arules/arules.pdf>
- Darwin, C. (1859). *Origin of the Species*.
- DataCamp. (2016). *Front page*. Retrieved from DataCamp:  
<https://www.datacamp.com/>



- Duke University. (2016). *Stationarity and differencing*. Retrieved from Duke University:  
<http://people.duke.edu/~rnau/411diff.htm>
- Entrepreneur. (2014). *Why Small Businesses Should Be Utilizing Customer-Loyalty Programs*.  
 Retrieved from Entrepreneur.
- Forecasting solutions. (2016). *Arima*. Retrieved from Forecasting solutions:  
<http://www.forecastingsolutions.com/arima.html>
- Frank, I. h. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*.  
 University of Waikato: Elsevier.
- Gartner. (2016). *Gartner Magic Quadrant*. Retrieved from Gartner:  
[http://www.gartner.com/technology/research/methodologies/research\\_mq.jsp](http://www.gartner.com/technology/research/methodologies/research_mq.jsp)
- Greiner, L. (2016, 01 7). *What is Data Analysis and Data Mining?* Retrieved 3 11, 2016,  
 from Database Trend and Applications:  
<http://www.dbta.com/Editorial/Trends-and-Applications/What-is-Data-Analysis-and-Data-Mining-73503.aspx>
- Han, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.
- High Need 2 Know. (2016). Retrieved from High Need 2 Know:  
<http://www.highneed2know.com/>
- Hu, S. (2007). *Akaike Information Criterion*. Retrieved from North Carolina State University: <http://www4.ncsu.edu/~shu3/Presentation/AIC.pdf>
- Inside-R. (2016). *strptime {base}*. Retrieved from Inside-R: <http://www.inside-r.org/r-doc/base/strptime>
- Investopedia. (2016, 03 11). *datamining*. Retrieved 03 11, 2016, from Investopedia:  
<http://www.investopedia.com/terms/d/datamining.asp>
- Investopedia. (2016). *Loyalty Program*. Retrieved from Investopedia:  
<http://www.investopedia.com/terms/l/loyalty-program.asp>
- IT Miner. (2015). *Data Mining Association Rule - Basic Concepts*. Retrieved from Youtube: <https://www.youtube.com/watch?v=RiFrbyiYpRs>
- Jinyan Zang and Cie. (2015). *Who Knows What About Me? A Survey of Behind the Scenes Personal Data Sharing to Third Parties by Mobile Apps*. Retrieved from Technology Science: <http://techscience.org/a/2015103001/>
- gybaudot. (2016, 19 01). *la tendance*. Retrieved 19 01, 2016, from gybaudot:  
<http://www.gybaudot.fr/Previsions/trend.html>

- KDnuggets. (2016, 03 03). *Gartner 2016 Magic Quadrant for Advanced Analytics Platforms: gainers and losers*. Retrieved 03 03, 2016, from KDnuggets.com:  
<http://www.kdnuggets.com/2016/02/gartner-2016-mq-analytics-platforms-gainers-losers.html>
- KDnuggets. (2016). *Gartner 2016 Magic Quadrant for Advanced Analytics Platforms: gainers and losers*. Retrieved from KDnuggets:  
<http://www.kdnuggets.com/2016/02/gartner-2016-mq-analytics-platforms-gainers-losers.html>
- Naisbitt, J. (1982). *Megatrends*. WARNER.
- Nestorov, S. (2000, 06 01). *DATA MINING TECHNIQUES FOR STRUCTURED AND SEMISTRUCTURED DATA*. Retrieved 03 11, 2016, from Stanford university: <http://infolab.stanford.edu/~evtimov/pubs/thesis.pdf>
- Oracle. (2016). *Descriptive Data Mining Models*. Retrieved from Oracle:  
[https://docs.oracle.com/cd/B12037\\_01/datamine.101/b10698/4descrip.htm](https://docs.oracle.com/cd/B12037_01/datamine.101/b10698/4descrip.htm)
- oText. (2016). *Time series components*. Retrieved 19 01, 2016, from oText:  
<https://www.otexts.org/fpp/6/1>
- Pollack, S. (2010). *Chapter 16: Time Series Analysis (1/4)*. Retrieved from Youtube:  
[https://www.youtube.com/watch?v=kJ\\_Os5iP0IA](https://www.youtube.com/watch?v=kJ_Os5iP0IA)
- probabilityislogic. (2011). *Practical thoughts on explanatory vs. predictive modeling*. Retrieved from Cross Validated:  
<https://stats.stackexchange.com/questions/1194/practical-thoughts-on-explanatory-vs-predictive-modeling>
- R Documentation. (2016, 18 03). Cross Tabulation and Table Creation. R *Documentation*.
- Rahm, E. (2016). *Data Cleaning: Problems and Current Approaches*. Leipzig: University of Leipzig, Germany.
- R-Bloggers. (2013, 07 21). *Read Excel files from R*. Retrieved 03 03, 2016, from <http://www.r-bloggers.com/read-excel-files-from-r/>
- Rithme. (2016). *Knowledge Discovery and Data Mining*. Retrieved from Rithme:  
<http://www.rithme.eu/?m=home&p=kdprocess&lang=en>
- Rstudio. (2014, 07 22). *Introducing tidy*. Retrieved 03 04, 2016, from Rstudio:  
<http://blog.rstudio.org/2014/07/22/introducing-tidy/>
- Rstudio. (2015, 08 31). *Introduction to dplyr*. Retrieved 03 03, 2016, from Rstudio:  
<https://cran.rstudio.com/web/packages/dplyr/vignettes/introduction.html>

Science Daily. (2016, 03 11). *Big Data, for better or worse: 90% of world's data generated over last two years*. Retrieved 03 11, 2016, from Science Daily:  
<https://www.sciencedaily.com/releases/2013/05/130522085217.htm>

Scientific Data. (2016). *Welcome to Scientific Data*. Retrieved from Nature:  
<http://www.nature.com/sdata/about>

Singhal. (2010, 01 27). *Time Series*. Retrieved 03 22, 2016, from Slideshare:  
<http://fr.slideshare.net/yush313/time-series-3000944>

Singhal, J. S. (2010, 01 27). *Time Series Analysis*. Retrieved 01 19, 2016, from Slideshare: <http://fr.slideshare.net/yush313/time-series-3000944>

Smart Vision. (2016, 03 07). *CRISP-DM stage three – data preparation*. Retrieved 03 07, 2016, from Smart Vision Europe: <http://www.sv-europe.com/data-preparation/>

Snowplow. (2016, 04 04). *Market basket analysis: identifying products and content that go well together*. Retrieved 04 04, 2016, from Snowplow:  
<http://snowplowanalytics.com/guides/recipes/catalog-analytics/market-basket-analysis-identifying-products-that-sell-well-together.html>

stackoverflow. (2016, 03 03). *tags*. Retrieved 03 03, 2016, from /stackoverflow:  
<https://stackoverflow.com/tags>

Statista. (2015). *Number of monthly active Facebook users worldwide as of 4th quarter 2015 (in millions)*. Retrieved from Statista:  
<http://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/>

Tableau Software. (2016). *A PROVEN LEADER*. Retrieved from Tableau:  
<http://get.tableau.com/gartner-magic-quadrant-2016.html>

Techaisle. (2015). *2015 Top 10 SMB Business Issues, IT Priorities, IT Challenges Infographic*. Retrieved from techaisle: <http://techaisle.com/smb-infographics/52-2015-top-10-smb-business-issues-it-priorities-it-challenges-infographic>

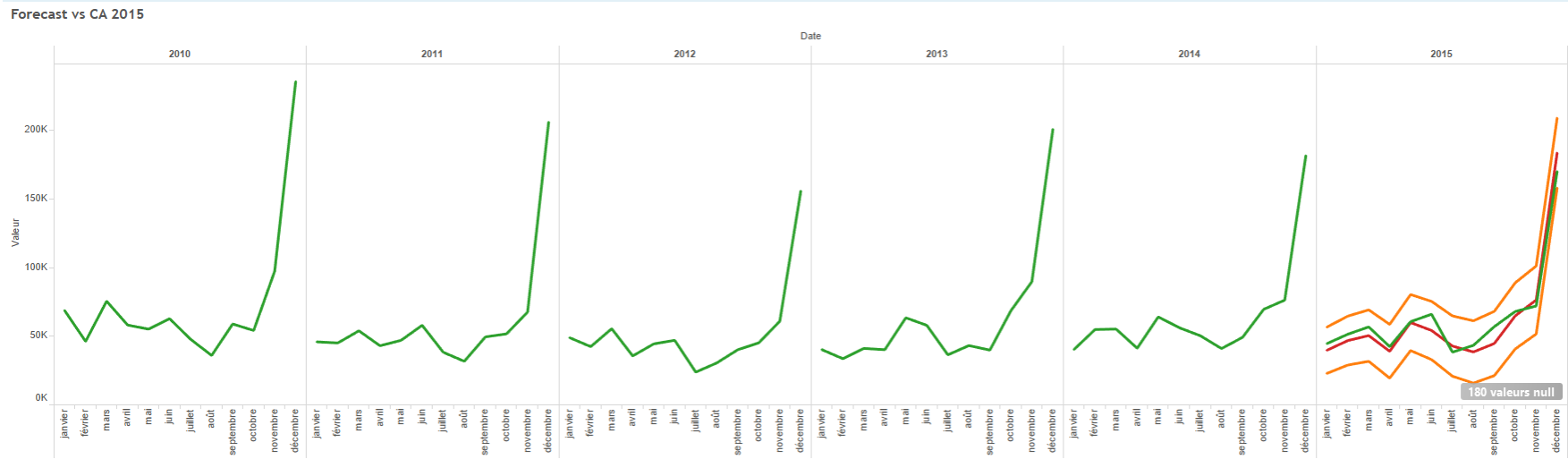
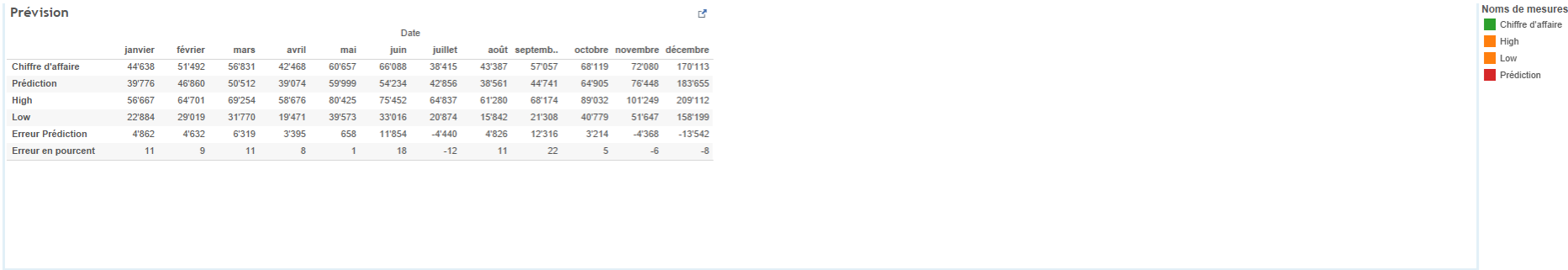
Technopedia. (2016). *Knowledge Discovery in Databases (KDD)*. Retrieved from Technopedia: <https://www.techopedia.com/definition/25827/knowledge-discovery-in-databases-kdd>

Technopedia. (2016, 03 11). *Techopedia explains Knowledge Discovery in Databases (KDD)*. Retrieved 03 11, 2016, from Technopedia:  
<https://www.techopedia.com/definition/25827/knowledge-discovery-in-databases-kdd>

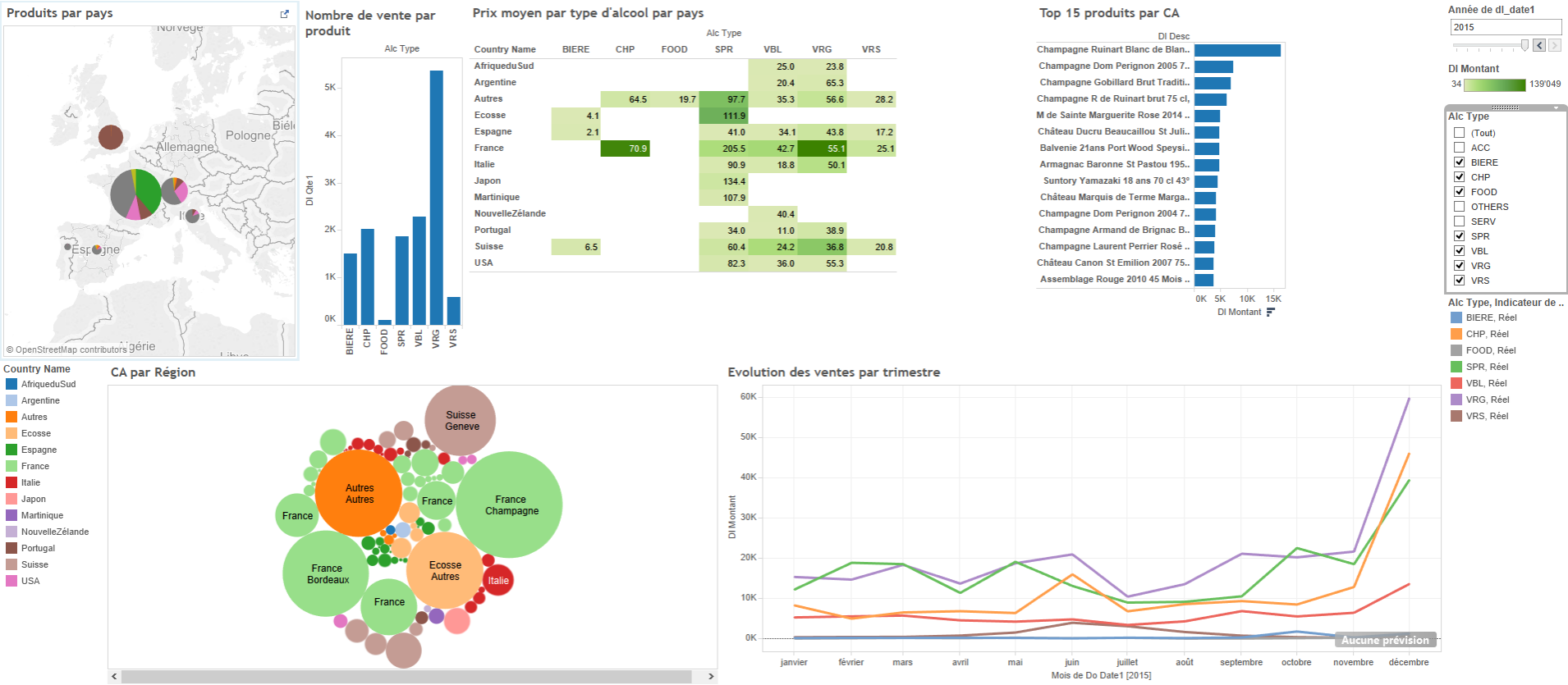
- The Register. (2006, 08 15). *The parable of the beer and diapers*. Retrieved 03 11, 2016, from The Register:  
[http://www.theregister.co.uk/2006/08/15/beer\\_diapers/](http://www.theregister.co.uk/2006/08/15/beer_diapers/)
- Trendowicz, A. (2012). *Software Cost Estimation, Benchmarking, and Risk Assessment*. Springer-Verlag Berlin and Heidelberg GmbH & Co.
- Tuffery, S. (2015). *Modélisation prédictive et apprentissage statistique avec R*. TECHNIP.
- Tutorials point. (2016). *Data Mining - Classification & Prediction*. Retrieved from Tutorials point:  
[http://www.tutorialspoint.com/data\\_mining/dm\\_classification\\_prediction.htm](http://www.tutorialspoint.com/data_mining/dm_classification_prediction.htm)
- UCLA Anderson. (2016). *Data Mining: What is Data Mining?* Retrieved from Anderson:  
<http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm>
- University of Minnesota. (2000, 02 10). *An Introduction to Cluster Analysis for Data Mining*. Retrieved 03 18, 2016, from University of Minnesota: An Introduction to Cluster Analysis for Data Mining
- vcefurthemaths. (2011, 03 24). *Maths Tutorial: Question on Data Transformations (statistics)*. Retrieved 03 18, 2016, from Youtube:  
<https://www.youtube.com/watch?v=EJ6EhfenqNs>
- Wikipedia. (2016, 03 03). *Missing data*. Retrieved 03 03, 2016, from Wikipedia:  
[https://en.wikipedia.org/wiki/Missing\\_data](https://en.wikipedia.org/wiki/Missing_data)
- Wikipédia. (2016, 19 01). *Série temporelle*. Retrieved 19 01, 2016, from Wikipédia:  
[https://en.wikipedia.org/wiki/Time\\_series](https://en.wikipedia.org/wiki/Time_series)
- Wikipédia. (2016, 03 18). *Smoothing*. Retrieved 03 18, 2016, from Wikipédia:  
<https://en.wikipedia.org/wiki/Smoothing>
- WinBIZ. (2016, 03 03). *Accueil*. Retrieved 03 03, 2016, from Winbiz.ch:  
<http://www.winbiz.ch/>
- Wpcurve. (2012). *Small Business Intelligence [INFOGRAPHIC]*. Retrieved from Wpcurve: <http://wpcurve.com/small-business-intelligence-infographic/>
- Zaïane, I. R. (1999). *Introduction to Data Mining*. University of Alberta, Computer Sciences.

Appendices

Appendix A: Tableau dashboard: 2016 Sales forecast per month



Appendix B: Tableau dashboard: Sales per type, country and region



Appendix C: Tableau dashboard: Margin analysis



Marge par Région

Région	DI Montant	DI Cle1	Prct Margin
Autres	220K	4.5K	51.29
Champagne	140K	2K	43.84
Geneve	60K	2K	46.83
Bordeaux	90K	1.5K	46.94
Bourgogne	40K	0.8K	46.81
Valais	20K	0.5K	50.75
Provence	10K	0.2K	45.91

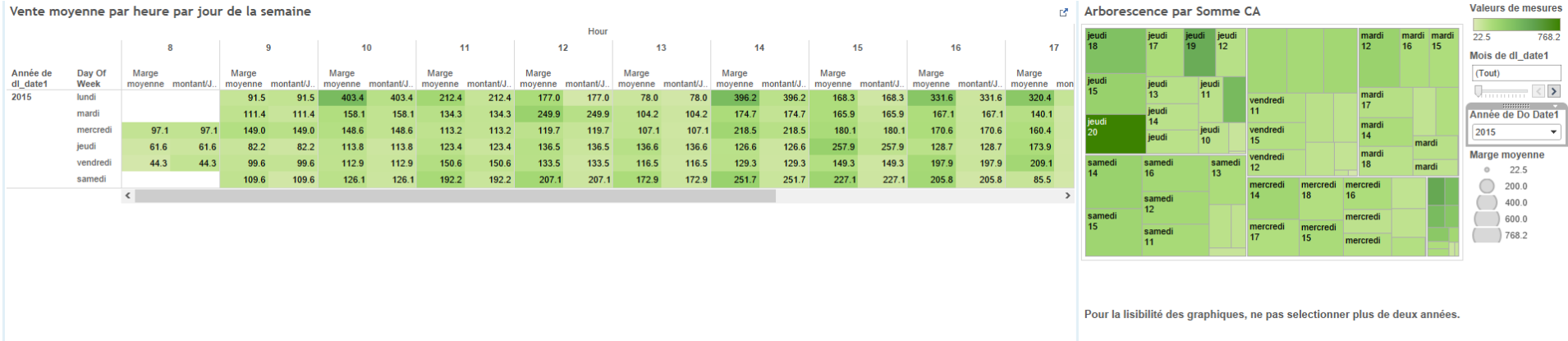
Marge par type d'alcool

Type d'alcool	DI Montant	DI Cle1	Prct Margin
VRG	250K	5.5K	47.75
VBL	70K	2.2K	47.18
CHP	140K	2K	43.81
SPR	200K	1.8K	47.46
BIERE	10K	0.5K	54.09

Marge par prix

Prix	DI Montant	DI Cle1	Prct Margin
(22,42)	140K	4.5K	53.061
(-90,22)	50K	4.2K	46.230
(42,79)	180K	3.2K	46.383
(79,159)	180K	2.2K	46.291

Appendix D: Tableau dashboard: Average sales per day of week and hours



CA par Heure

